Population-wise Consistent Segmentation of Diffusion Weighted Magnetic Resonance Images

by

Ulas Ziyan

B.S. Electrical Engineering, Purdue University, 2000
M.S. Electrical and Computer Engineering, Purdue University, 2001

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2008

© Massachusetts Institute of Technology 2008. All rights reserved.

Author .....
Department of Electrical Engineering and Computer Science
May 23, 2008

Certified by .....
Carl-Fredrik Westin
Associate Professor of Radiology at Harvard Medical School
Thesis Supervisor

Certified by .....
W. Eric L. Grimson
Bernard Gordon Professor of Medical Engineering
Thesis Supervisor

Accepted by .....
Terry P. Orlando
Chairman, Department Committee on Graduate Students

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
JUL 0 1 2008
LIBRARIES

ARCHIVES

Population-wise Consistent Segmentation of Diffusion Weighted Magnetic Resonance Images

by

Ulas Ziyan

Submitted to the Department of Electrical Engineering and Computer Science
on May 23, 2008, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

In this thesis, we investigate unsupervised and semi-supervised methods to construct anatomical atlases and segment medical images. We propose an integrated registration and clustering algorithm to compute an anatomical atlas of fiber-bundles as well as deep gray matter structures from a population of diffusion tensor MR images (DT-MRI).

We refer to this algorithm as "Consistency Clustering" since the outputs of the algorithm include population-wise consistent segmentations and correspondence between the subjects. The consistency is ensured through using a single anatomical model for the whole population, which is similar to the atlases used by experts for manual labeling. We experiment with both parametric and non-parametric models for the gray matter and white matter segmentation problems, each model resulting in a different kind of atlas.

Consistent population-wise segmentations require development of several integrated algorithms for clustering, registration, atlas-building and outlier rejection. In this thesis we develop, implement and evaluate these tools individually and together as a population-wise segmentation tool. Together, Consistency Clustering enables automatic atlas construction in DT-MRI for a population, either normal or affected by a neural disorder. Consistency Clustering also provides the user the choice to include prior knowledge through a few labeled subjects (semi-supervised) or compute an anatomical atlas in a completely data-driven manner (unsupervised). Furthermore, resulting anatomical models are compact representations of populations and can be used for population-wise morphometry.

We implement and evaluate these methods using in vivo DT-MRI datasets. We investigate the benefits of population-wise segmentation as opposed to individually segmenting subjects, as well as effects of noise and initialization on the segmentations.

Thesis Supervisor: Carl-Fredrik Westin
Title: Associate Professor of Radiology at Harvard Medical School

Thesis Supervisor: W. Eric L. Grimson
Title: Bernard Gordon Professor of Medical Engineering

Acknowledgments

Three months ago when Eric told me that it was time to start writing the thesis, I was very excited that this is finally coming to an end. Now, in these final hours before the thesis deadline, I am a little sad that the end is so near. Over the last couple years, I became used to the PhD student lifestyle at MIT and I enjoyed a seemingly infinite intellectual freedom. But, I do not think I appreciated neither one as much as I should have. The PhD experience was challenging (to put it lightly) but I am glad I did it. I think I now am a better person for it, not just due to the academic challenge of the program, but more importantly because of the great people I worked with.

I am grateful to David Tuch, who introduced me to diffusion tensor imaging and help me secure the Martinos Center funding. It was Carl-Fredrik; however, whose research interests matched with mine. He served as a great mentor and our relationship felt more of a friendship than just a professional one. Eric was always there when I needed him, even though he progressively got busier as he took on additional responsibilities with the department. Every time I talked with him, he made me feel good about what I was doing. Without his constant support, I don't think I would be able to reach the finish line.

I would like to thank both Alan Willsky and Elfar Adalsteinsson for serving in my thesis committee and for their encouragement and constructive feedback. I thank Mert Sabuncu for introducing me to the vast field of registration, and Lauren O'Donnell for sharing with me her white matter clustering code and results. Mert, Wanmei, Thomas, Serdar and Bis all read some parts of this document and gave great feedback. I appreciate their time and honesty.

Of course, I thank my parents, who constantly surprise me with their seemingly infinite wisdom. A simple phone conversation with them would put everything into perspective, even during the worst of times. Also, I thank my brother, Ahmet, for his constant encouragement.

Finally, I would like to thank Cynthia, the person who is dearest to me. Thanks Cynthia, for being yourself, being there and supporting me at every step of the way.

Contents

1 Introduction **17**

1.1 Existing Segmentation Methods for DTI 18

1.1.1 Tensor similarity based segmentation 19

1.1.2 Connectivity based segmentation 19

1.2 Problems with existing methods 21

1.3 Objectives and Contributions of the Thesis 22

1.4 Organization of the Thesis 23

2 Human Brain **27**

2.1 Introduction 28

2.2 White Matter 30

2.3 Deep Gray Matter 32

2.4 Summary of the Chapter 35

3 Diffusion Weighted Magnetic Resonance Imaging **37**

3.1 Diffusion 38

3.2 Quantifying Diffusion: Pulse Gradient Spin Echo 38

3.3 Diffusion in Free Water 41

3.4 Diffusion in Tissue 42

3.5 Diffusion Tensor Imaging 43

3.5.1 Diffusion Tensor Estimation 44

3.5.2 Eigensystem of Diffusion Tensor 44

3.5.3 Scalar Derivatives of Diffusion Tensor 45

3.6 Higher Order Diffusion Models 47
3.7 Tractography 47
3.8 Summary of the Chapter 48

4 Unsupervised Clustering as a Deep Gray Matter Segmentation Tool 49

4.1 Introduction 50
4.2 Background 53
4.2.1 k-means 53
4.2.2 Spectral Clustering 56
4.3 Spectral Split & Merge Clustering 61
4.4 Application of Spectral Split & Merge Clustering to DTI 67
4.4.1 Graph Construction 67
4.4.2 Graph Cuts 69
4.5 Data Collection and Pre-Processing 70
4.5.1 Image acquisition 70
4.5.2 Pre-processing 70
4.5.3 Thalamus Mask 71
4.5.4 Manual Nuclei Segmentation 72
4.5.5 Validation 73
4.6 Experiments 75
4.6.1 Comparison of Heuristic Solutions to Normalized Cuts 75
4.6.2 Comparison of Tensor Similarity Measures 75
4.6.3 Qualitative Results from Spectral Clustering on DTI Data 76
4.6.4 Effect of k 77
4.7 Summary of the Chapter 83

5 Consistent Gray Matter Segmentation 85

5.1 Introduction 86
5.2 Background on Expectation Maximization 89
5.2.1 An Iterative Lower Bound Optimization 89
5.2.2 Maximum Likelihood Density Estimation 89

5.2.3 EM for Estimating a Gaussian Mixture Model 94
5.2.4 EM for Estimating a von-Mises Fisher Mixture Model 98
5.3 Consistency Clustering 103
5.3.1 E-Step 104
5.3.2 M-Step 105
5.3.3 Prior Information 108
5.3.4 Implementation Details 110
5.4 Experiments 113
5.4.1 Data Collection and Pre-Processing 113
5.4.2 Individual versus Population-wise Segmentation 113
5.4.3 Effect of Varying Amounts of Prior Information 117
5.4.4 Comparison of Segmentation Methods 124
5.5 Summary of the Chapter 128

6 Consistent White Matter Segmentation 131
6.1 Introduction 132
6.2 Data Collection and Pre-Processing 136
6.2.1 Image acquisition 136
6.2.2 Tractography 136
6.2.3 Initial Registration 137
6.2.4 Fiber Clustering for a High Dimensional Atlas 137
6.3 Consistency Clustering for White Matter Segmentation 139
6.3.1 E-Step 141
6.3.2 M-Step 142
6.3.3 Tract Cut 143
6.3.4 Outlier Rejection 144
6.4 Experiments 146
6.4.1 Consistency Clustering: A Clustering Algorithm 147
6.4.2 Consistency Clustering: An Atlas-building Algorithm 147
6.4.3 Atlas Quality 150

6.5 Summary of the Chapter 155

7 Conclusion **157**

Bibliography **161**

List of Figures

1-1 Examples of tractography and clustering errors 20

1-2 Correspondence between clusters in different subjects 21

1-3 Schematic description of Consistency Clustering 22

1-4 Effect of outlier rejection and registration on atlas 23

1-5 Effect of prior information on segmentation results 24

1-6 Example white matter segmentation obtained through consistency clustering 25

2-1 Structure of a typical neuron 28

2-2 Lobes of the cerebral cortex 29

2-3 Projection tracts 31

2-4 Association tracts 32

2-5 Corpus callosum 33

2-6 Cingulum 33

2-7 Fornix 34

2-8 Thalamus 34

3-1 Pulse sequence diagram for PGSE 39

3-2 Illustration of anisotropic diffusion in fibrous tissue 42

3-3 Illustration of diffusion tensor in fibrous tissue 45

3-4 Common scalar derivatives of the diffusion tensor 46

4-1 Three-dimensional DTI rendering of a thalamus 51

4-2 Inter-subject variability in thalamus 52

4-3 A toy example to illustrate reduction in NCut through our spectral algorithm 62

4-4 Qualitative comparison of heuristic solutions to normalized cuts in DTI segmentation 76
4-5 Quantitative comparison of heuristic solutions to normalized cuts in DTI segmentation 77
4-6 Comparison of similarity measures in DTI segmentation with $k = 7$ 78
4-7 Comparison of similarity measures in DTI segmentation with $k = 12$ 78
4-8 A schematic outline of spectral segmentation algorithm 79
4-9 3D rendering of thalamic nuclei after spectral segmentation 80
4-10 An example hierarchy tree obtained through spectral clustering 81
4-11 Effect of increasing k on volume overlap measure 82

5-1 Differences between previous thalamus segmentation algorithms and Consistency Clustering 87
5-2 Illustration of an EM iteration 90
5-3 Individually segmented thalamic nuclei 114
5-4 Simultaneously segmented thalamic nuclei 115
5-5 Quantitative comparison of individual versus population-wise segmentation 116
5-6 Results from CC with seven labeled voxels used as prior information. . . . 118
5-7 Results from CC with single labeled subject used as prior information . . . 119
5-8 Impact of the choice of labeled subject on the segmentation results 120
5-9 Results from CC with nine labeled subject used as prior information 121
5-10 Expert Labeled Thalami 122
5-11 Impact of varying amounts of prior information on the segmentation results 123
5-12 Segmentation results from k-means algorithm 125
5-13 Segmentation results from spectral clustering algorithm 126
5-14 Volume overlaps from different segmentation methods 127

6-1 Some common tractography and clustering errors 134
6-2 Illustration of the tract cut operation 144
6-3 Illustration of outlier removal 145
6-4 Consistency clustering results from a synthetic dataset 148

6-5 Stability of labels produced by Consistency Clustering 149
6-6 White matter atlas results 151
6-7 Fiber bundle cores identified through Consistency Clustering 152
6-8 Consistency of atlas-based labeling 153
6-9 3D fiber bundle model 154
6-10 Impact of varying outlier class probability level 154

List of Tables

4.1 Thalamic nuclei labels 72

6.1 Experimental atlas entropies from different methods 150

Chapter 1

Introduction

Human brain is the control center of the human body, responsible for virtually all human activity, including movement, thought and emotions. It does so by receiving and processing external sensory information from other parts of the body, and sending electrical pulses back to the body through the spinal cord and the peripheral nervous system. While it is not completely understood how the brain processes this information, neuroscientists have uncovered considerable knowledge of the brain's anatomy.

The human brain is not a single homogeneous entity, rather it consists of hundreds of distinct functional structures in the form of gray matter, and numerous connections between these structures in the form of white matter tracts. The field of neuroscience is often interested in studying the effects of neural diseases on specific structures. Some of these effects could be a change in size, shape or tissue integrity. For example, it has been reported that in the case of schizophrenia, the size of thalamus, an organ in the deep gray matter, decreases [22] and the white matter connectivity in the fornix, corpus callosum, and cingulum tracts degenerates [39].

These studies usually depend on manual segmentations obtained by trained experts from structural magnetic resonance (MR) images. However, structural MRI offers little or no contrast in many areas of interest, such as in deep gray matter or along white matter tracts. This is why the recent introduction of diffusion weighted imaging (DWI) as a new MRI modality has caused much excitement.

DWI is a relatively new imaging modality that measures free water diffusion, i.e. Brow-

nian motion, of the endogenous water in tissue [6]. This water diffusion is measured for the entire brain using directional gradients in the MRI machine. Unlike structural MRI, however, DWI does not result in a single image, but a series of images, each one quantifying the amount of water diffusion in a specific direction. These images are then combined together to calculate a 3D water diffusion probability distribution function (pdf) for every voxel. The most common way of calculating this pdf is to assume the water diffusion is Gaussian for a given voxel and fit the data from individual DWI images to this model, resulting in diffusion tensor images (DTI).

In human brain tissue, the water diffusion is not the same in all directions, since it is obstructed by structural elements such as cell membranes or myelin [6]. When this obstruction constrains the water diffusion in a coherent direction, such as within the cerebral white matter, the resulting water diffusion tensor becomes anisotropic, containing information about the directionality of the axon bundles. Thus, quantification of water diffusion in tissue through DTI provides a unique way to look into the organization of the brain.

Unlike white matter, the tissue in gray matter is less organized by orientation. The lack of coherent orientation limits the use of DTI for gray matter analysis in some areas, such as the cerebral cortex. However, there are certain gray matter structures, such as the thalamus, that exhibit coherence in diffusion direction due to the presence of coherent white matter near these structures. Thalamus and several other deep gray matter structures are organized into distinct functional regions, called nuclei. Functionally related white matter tracts target the same region of cortex once they leave the deep gray matter, resulting in organization of diffusivity within the gray matter. This organized diffusion can be measured in DTI, and it has been proposed that the thalamic nuclei can be distinguished by their characteristic diffusion orientation [81].

1.1 Existing Segmentation Methods for DTI

Since the realization that both white and gray matter can be resolved through DTI, several segmentation algorithms have been proposed to segment individual structures using DTI. Previous research work on DTI segmentation can be categorized into two branches: Tensor

similarity based methods and connectivity based methods.

1.1.1 Tensor similarity based segmentation

The traditional way to obtain an image segmentation is to define a similarity measure between pixels (voxels in 3D images) and then to calculate a solution that maximizes a segmentation quality measure that depends on this similarity. The most commonly used similarity measure in the DTI literature is the Frobenius norm between two tensors [81, 21, 65, 73]. There are also other measures that depend on fractional anisotropy (a scalar invariant of the diffusion) [85], symmetrized Kullback-Leibler divergence between tensors [74], normalized tensor scalar product [31, 32], and angular distance between principle diffusion directions [60]. Once a tensor similarity is chosen, the next step is to identify a quality measure and an optimizer. Different segmentation methods are proposed in the literature, including k-means algorithm [81], mean-shift algorithm [19], region-based and edge-based level sets [85, 21, 65, 73, 31]. Most of these methods require a good initialization for successful segmentation. The initialization methods used were atlas-based [81, 32], or manual [85, 21, 73, 44, 74]. These methods were demonstrated both on gray [81, 32, 19] and white matter segmentations [85, 21, 31, 73, 65, 44, 74].

1.1.2 Connectivity based segmentation

Since DTI contains information about the directionality of the water diffusion, it is interesting to attempt to reconstruct the underlying white matter tracts. Reconstruction typically consists of starting at a voxel and following the most likely path of diffusion until a stopping criteria is reached. This sort of reconstruction is called tractography and results in a set of 3D curves, called fibers. If tractography is seeded at every white matter voxel in the brain, the resulting set of fibers span the whole brain. We know that typical in-vivo imaging resolution is too low to reconstruct individual axons from the DTI images, but since these axons tend to move coherently in space, the larger white matter tracts appear as fiber bundles in tractography.

Since larger white matter tracts become evident as fiber bundles in tractography, a natu-

Figure 1-1: Three segmented fiber bundles from two subjects. Corpus callosum is shown in black; cingulum is in purple and fornix is in green. A: Example of where the tractography deviated from cingulum bundle to corpus callosum. B: Example of where tractography failed to track the whole length of the cingulum bundle.

ral extension of tractography is to identify these bundles and assign them anatomical labels. A popular technique to assign white matter anatomy to the fiber bundles is to manually select regions of interests (ROI) that are thought to correspond to a particular anatomical white matter tract, and label all the fibers passing through this ROI with the corresponding anatomical label [83]. In several studies it was reported that using two ROIs per white matter tract results in more stable fiber bundle assignments than using only one ROI [28].

ROI based methods could be subject to user bias if the regions of interests are manually traced. Therefore several methods have been proposed to identify anatomically meaningful regions from the DTI data. One class of methods uses fibers generated via tractography and groups them into regions either interactively or automatically [46, 47, 56].

The connectivity information is also useful in segmenting the deep gray matter nuclei. Behrens et al. utilized probabilistic tractography, which is seeded at every voxel in the thalamus. Since the white matter pathways connect the functional areas of thalamus to the functionally related areas of the cortex, this method results in a probabilistic map from every voxel in the thalamus to the cortex. They also hand segmented the cortex into its functional areas, and used the probabilistic mapping from the cortex to the thalamus to infer a segmentation of the thalamus.

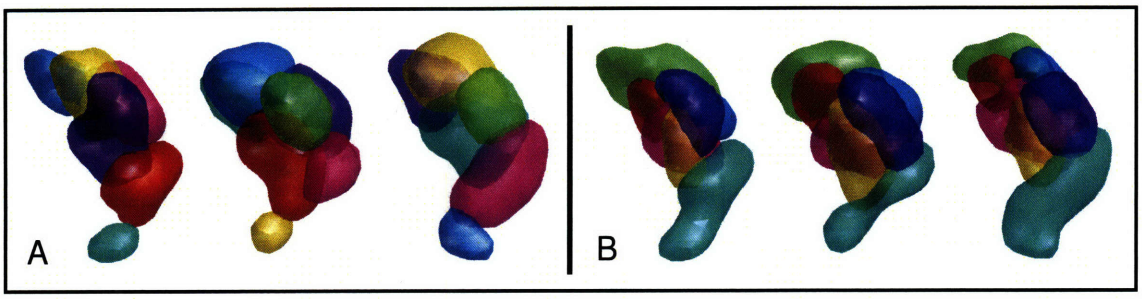

Figure 1-2: Thalamus data from three subject is segmented individually using the k-means algorithm (A). The resulting segmentations do not show correspondence among the subjects as indicated by the coloring of the segmentations. In this thesis we propose a segmentation algorithm that ensures consistency, as shown in (B). The consistency is ensured by segmenting a population simultaneously.

1.2 Problems with existing methods

DTI suffers from low SNR and resolution, even more so than a regular MRI because of the need for collecting many images to construct one DTI data set. This fact becomes apparent in tractography and usually results in fibers deviating from one tract bundle to another and fibers not fully covering the entire length of the bundle [83]. (See Figure 1-1 for effects of noise on tractography).

Therefore, it is not adequate to cluster fibers into bundles and assign them anatomical labels; there is also a need to ensure that the tractography noise is taken into account and remedied before being able to segment the underlying DTI data. To our knowledge, this is an open problem, and an important one since being able to segment the white matter anatomy in a precise manner is the first step of quantitative analysis of white matter.

Furthermore, even though most of the tensor similarity based segmentation methods summarized in Section 1.1.1 produce plausible segmentations for any given subject in a population, they do not find a correspondence between the segments acquired from different subjects (Figure 1-2). Without correspondence between segmentation results, it is not possible to assign consistent anatomical labels to individuals.

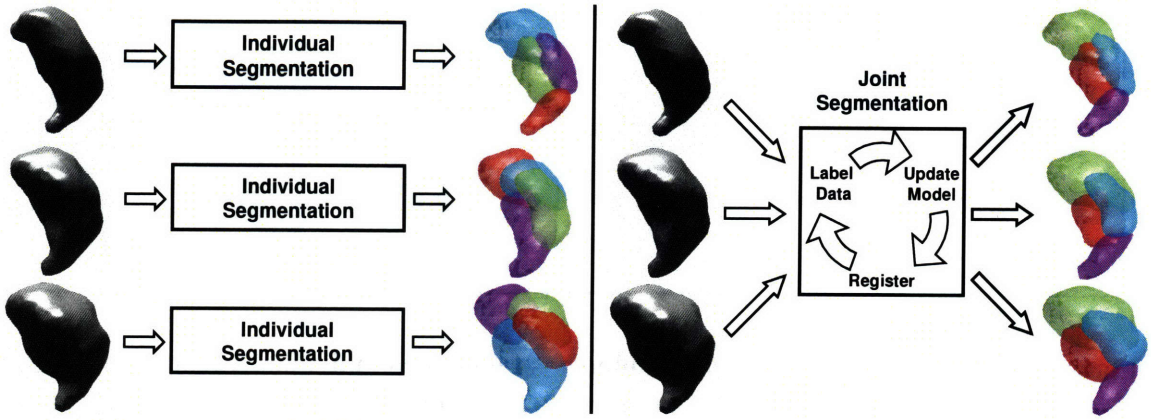

Figure 1-3: Schematic description of previous thalamus segmentation algorithms [81, 86, 32, 19] (left) as opposed to the Consistency Clustering algorithm (right).

1.3 Objectives and Contributions of the Thesis

The main objective of this thesis is to derive, implement and evaluate a population-wise DTI segmentation algorithm for white matter tracts as well as gray matter nuclei. We refer to this algorithm as "Consistency Clustering" (CC) since the outputs of the algorithm are population-wise consistent segmentations, as well as a correspondence between the subjects (Figure 1-3). The consistency is ensured through using a single anatomical model for the whole population, which is similar to the atlases used by experts for manual labeling. We experiment with both parametric and non-parametric models for the gray matter and white matter segmentation problems, each model resulting in different kinds of atlases.

The Consistency Clustering algorithm requires development of several integrated algorithms for segmentation, registration, atlas-building and outlier rejection. The segmentation is performed by labeling each voxel or fiber in the dataset using a probabilistic model. The probabilistic model served as an anatomical atlas, and it is learned iteratively from labeled data while performing the segmentation. To reduces the effect of noise on the learned model, an outlier rejection is performed by introducing a default label with low probability level everywhere in the model (see Figure 1-4 for effect of outlier rejection on the learned atlas.) Finally, the registration (spatial normalization) is achieved through a poly-affine framework, that consist of a single affine warp per cluster in the model.

In this thesis we develop, implement and evaluate these tools individually and together

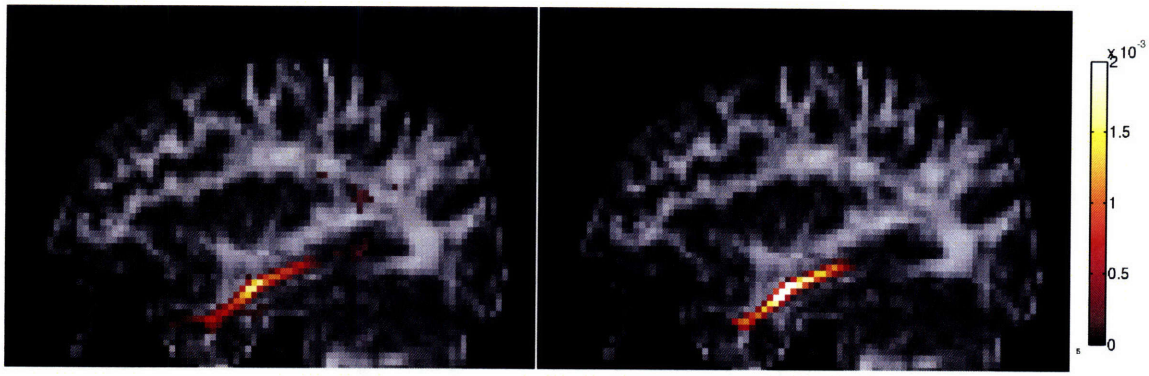

Figure 1-4: White matter atlases learned from 15 subjects' tractography results. The left image depicts the spatial distribution of fornix fibers segmented using spectral clustering without outlier rejection. The right image shows the spatial distribution of the same fibers segmented with Consistency Clustering, which includes an outlier rejection scheme. Consistency clustering removes dispersed low probability areas (dark red speckles) in the atlas and assigns more of the probability mass to the core of the fornix, which is apparent by higher amounts of white color in the atlas.

as a population-wise segmentation tool. Together, the Consistency Clustering enables automatic atlas construction in DTI for a population, either normal or affected by a neural disorder. Consistency Clustering also provides the user the choice to include prior knowledge through a few labeled subjects (semi-supervised) or to compute an anatomical atlas in a completely data-driven manner (unsupervised) (See Figure 1-5 for example segmentations obtained either way). Consistency Clustering is implemented and evaluated on deep gray matter (Figure 1-5) and also on white matter tracts (Figure 1-6). In either case, the resulting anatomical models are compact representations of populations and can be used for population-wise morphometry.

1.4 Organization of the Thesis

In Chapters 2 and 3, we provide brief background information on the human brain and on diffusion tensor imaging, to familiarize the reader with these subjects. In Chapter 4, we experiment with several unsupervised clustering algorithms for the deep gray matter segmentation problem. The chapter highlights the shortcomings of individual clustering of subjects as a segmentation tool, and provides comparisons between different tensor similarity metrics for DTI segmentation. In Chapters 5 and 6, we formalize the idea of consistent

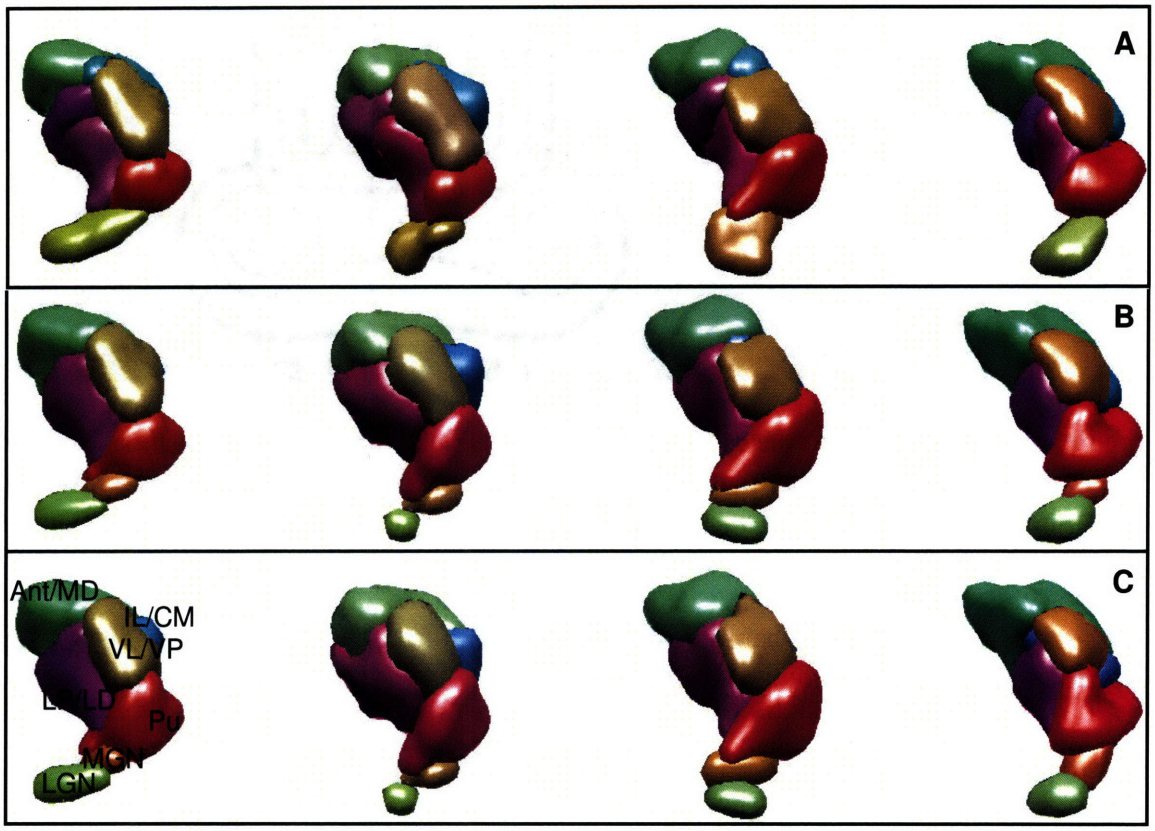

Figure 1-5: Segmentations of four thalami obtained through (A) Consistency Clustering without prior information (unsupervised), (B) Consistency Clustering with few labeled subjects used as prior information (semi-supervised), and (C) expert labeling. Consistency Clustering results in consistent segmentations among the subjects with or without prior information. However, prior information results in segmentations more similar to the expert labeled thalami (ventral nuclei, purple nuclei on the left side of the images, are divided into two pieces without prior information, whereas they are not divided when segmented with prior information.)

clustering and derive specific equations for the algorithm for deep gray matter and white matter tract segmentation problems. We also provide experiments investigating the benefits of population-wise segmentation. Finally, in Chapter 7, we summarize our findings and provide a discussion for future research.

Figure 1-6: Tracts from Fornix (in green) and Cingulum (in purple) bundles along with a few selected tracts from Corpus Callosum (in black) as labeled through spectral clustering (left) and through Consistency Clustering (right). The tractography noise is evident in the images on the left as tracts deviating from one bundle to another. Also, these images contain instances where the high dimensional atlas failed to label the tracts correctly. Consistency Clustering is able identify consistent tract bundles across subjects, while removing the inconsistent parts of these bundles, resulting in "core" bundles.

Chapter 2

Human Brain

The purpose of this chapter is to familiarize the reader with human brain anatomy, particularly the anatomy of the white matter tracts and deep gray matter, whose automatic segmentation is the topic of the thesis.

We begin with a review of the human brain tissue, which consists of two distinct regions, gray matter and white matter. We then investigate the composition and organization of the white matter and deep gray matter. While providing a basic background on brain anatomy, we also highlight several structures that are investigated in the thesis.

The human brain is a complex structure that fascinates many, and attracts scientists from all backgrounds. It is not practical to include a complete survey of this large set of studies in this chapter; however, we recommend [55, 25, 82] to the interested reader for a deeper understanding of the human brain anatomy.

2.1 Introduction

The human brain is a portion of the central nervous system that is located within the skull. It is the control center of the human body and it is responsible for virtually all human activity, including movement, thought and emotions. It operates by receiving and processing external sensory information from other parts of the body, and sending electrical pulses back to the body through the spinal cord and the peripheral nervous system. It is not completely understood how the brain processes this information, however, neuroscientists have been studying the anatomy of the brain for hundreds of years and through this research we learned quite a lot about the brain's anatomy.

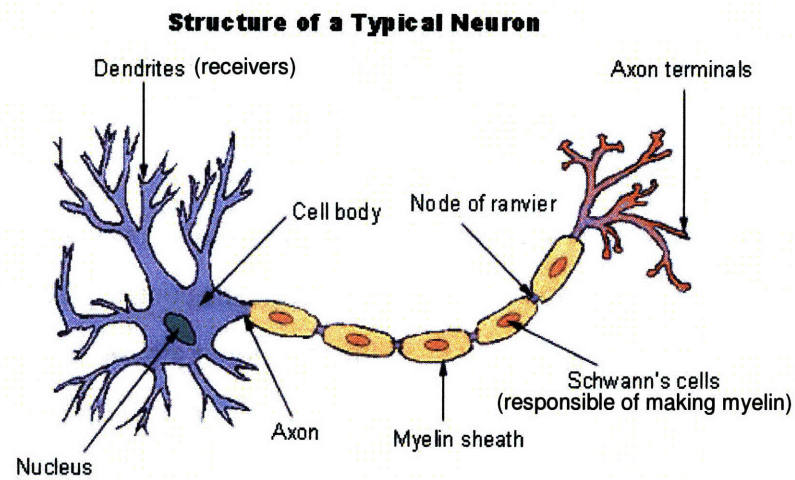

Figure 2-1: Structure of a typical neuron. A typical neuron consist of dendrites to receive information, a cell body to process the information and an axon to transmit information to other cells. Image is adapted from [84].

At the microscopic scale, the anatomy of brain consists of neurons and glial cells. The glial cells account for a majority of the brain mass and together they provide physical and physiological support to the neurons. Neurons are nerve cells and they are responsible for receiving, processing and transmitting information as electrical pulses throughout the brain. A typical neuron consist of dendrites to receive information, a cell body to process the information and an axon to transmit information to other cells. An illustration of a typical neuron is presented in Figure 2-1. When an electrical impulse is received at one of the dendrites, the neuron may depolarize, transmitting the electrical signal through the

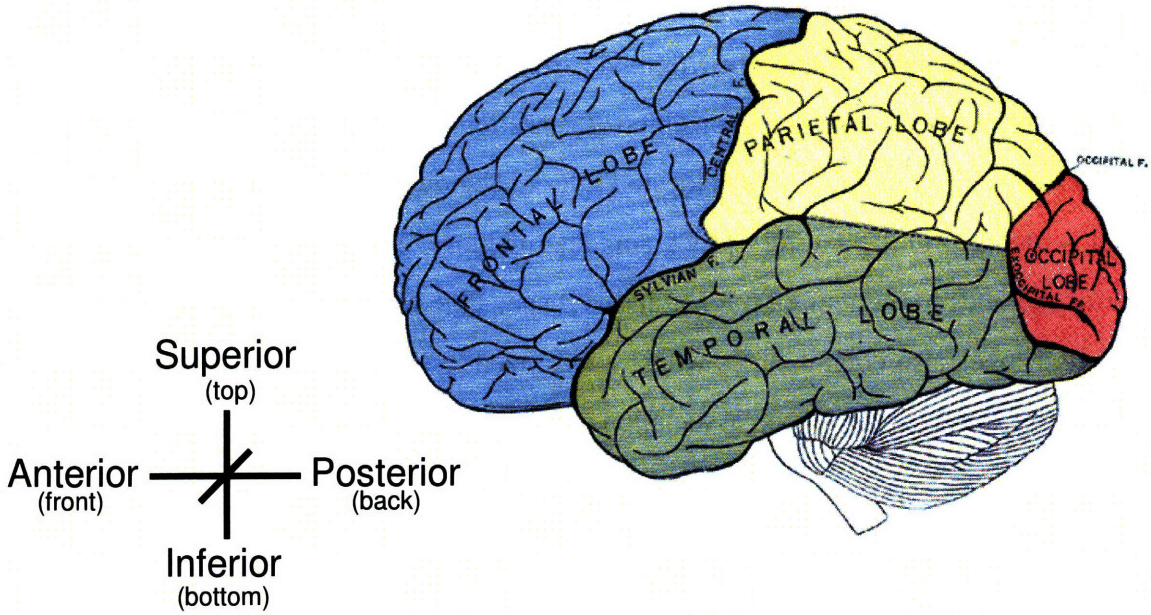

Figure 2-2: The four lobes of the cerebral cortex along with a sketch illustrating the orientation of the image. Image is adapted from [25].

length of the axon to the next neurons. The axons are covered with a white substance called myelin, which acts as an insulator, increasing the speed of the electrical signal transmission.

At a larger scale, the brain consist of cerebrospinal fluid (CSF), gray matter, and white matter. The CSF fills the space in the ventricular system around and inside the brain, providing physical support and immunological protection to the rest of the brain. The gray matter consists mainly of the neural cell body, which is gray in color. The majority of the gray matter surrounds the perimeter of the brain, following a highly convoluted manifold, which is called the cerebral cortex. The cerebral cortex is responsible for the majority of the cognitive functions. A smaller portion of the gray matter is located deep inside the brain, at the other end of the neurons, called the deep gray matter. White matter consists of the myelinated axons, interconnecting the gray matter. Its name and color is due to the white color of the myelin sheath that cover the axons.

At the macroscopic scale, the brain consists of two hemispheres, which are largely seperated by a deep groove, but connected by a thick bundle of neurons, called the corpus callosum (Figure 2-5). Each hemisphere of the cerebral cortex can be separated into four functional lobes (Figure 2-2). The frontal lobe includes the motor cortex and has a major

role in planning and execution of movements. The central sulcus (a prominent valley in the cortex) separates the frontal lobe from the parietal lobe. The parietal lobe contains the primary sensory cortex and is responsible for sensations, spatial orientation and information processing. The temporal lobe is located inferior (below) to the frontal and parietal lobes, and is involved in the auditory processing and memory. The occipital lobe is located posterior (behind) to the temporal lobe, and is responsible for processing visual information. All four lobes also contain areas whose specialized function have not yet been identified. These areas are known as the association cortex and are subject of future neuroscientific research.

2.2 White Matter

White matter consists of myelin covered axons, which are involved in transmitting electrical pulses from one functional region of the brain to another. The myelin provides electrical insulation for faster transmission, and its white color is the reason for the name white matter.

The axons can either occur in a diffusive manner or concentrated in bundles, forming white matter fiber tracts. The white matter contains three different types of neural tracts, which are called commissurral tracts, projection tracts and association tracts.

Commissurral tracts are white matter fiber tracts that connect functionally related regions in one hemisphere to related regions in the other hemisphere. Projection tracts establish connections between the cerebral cortex and the sub-cortical gray matter structures, such as the thalamus (Figure 2-3). Virtually all sensory information and motor commands travel through the projection tracts to and from the cerebral cortex. And finally the association tracts are shorter tracts near the cerebral cortex, connecting one part of the cortex to another. Association tracts come in different lengths, connecting two regions within the same gyrus, connecting two gyri or connecting two functional lobes (Figure 2-4).

We now highlight some of the white matter tracts that are investigated in the thesis. We investigate these tracts, because they are major fiber tracts in close proximity to each other, and this proximity offers challenges to a segmentation algorithm.

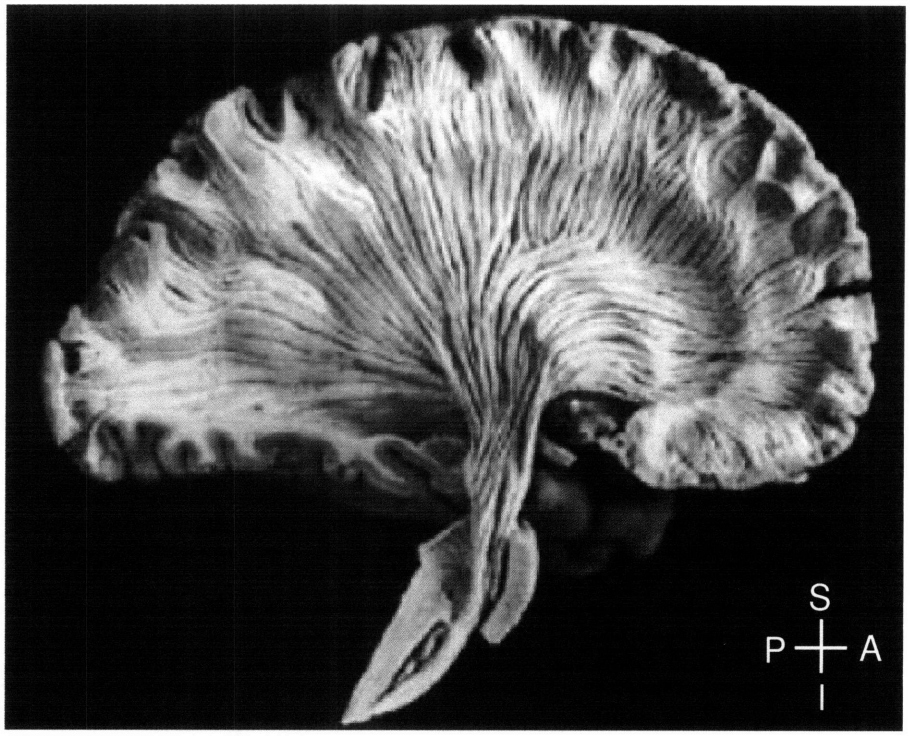

Figure 2-3: Projection tracts from a dissected brain. Projection tracts establish connections between the cerebral cortex and the sub-cortical gray matter structures. Image is adapted from [82].

Corpus Callosum

The corpus callosum is a commissurral tract bundle that connects the two hemispheres of the brain. It is the largest fiber bundle in the brain, and accounts for most of the inter-hemispheric communication. It forms a concentrated neural bridge between the hemispheres near the middle of the brain and diverges to a wide area closer to the cortex (Figure 2-5).

Cingulum

The cingulum is a prominent association tract bundle that is located adjacent but superior to the corpus callosum. The axons in the cingulum are oriented in the anterior-posterior direction as opposed to the lateral corpus callosum fibers (Figure 2-6). The cingulum is part of the limbic system, which also includes the hippocampus and amygdala, and supports a variety of functions including emotion, behavior and long term memory.

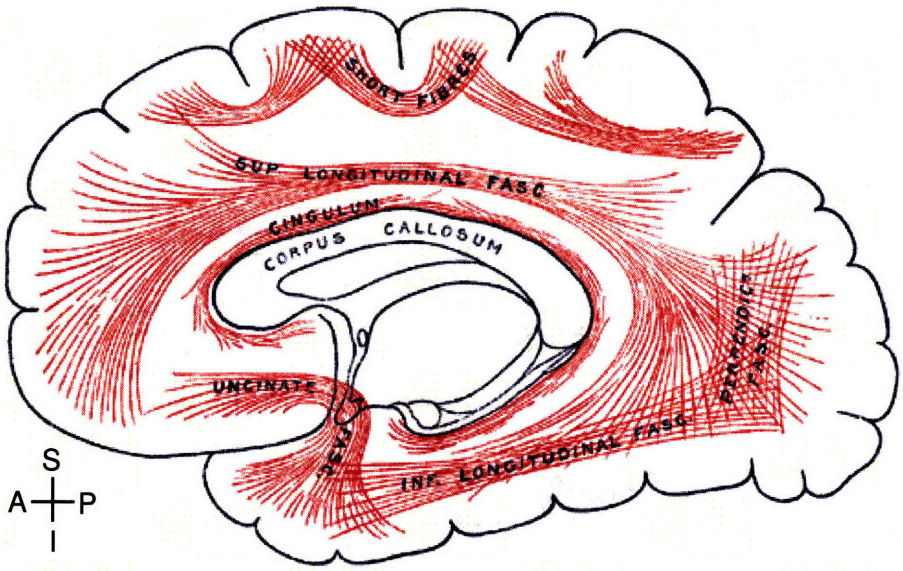

Figure 2-4: Illustration of association tracts. Association tracts are located near the cerebral cortex, connecting one part of the cortex to another. Association tracts come in different lengths, connecting two regions within the same gyrus, connecting two gyri or connecting two functional lobes. Image is adapted from [25].

Fornix

The fornix is also part of the limbic system, connecting hippocampus to mammillary bodies. The latter are deep gray matter nuclei, and act as relay stations for impulses that originate at the hippocampus. Fornix is located inferior to the corpus callosum and travel in the anterior-posterior direction, similar to cingulum (Figure 2-7).

2.3 Deep Gray Matter

Gray matter contains cell bodies and dendrites of neurons, and exhibits a gray brown color due to blood vessels and the neural cell bodies. The majority of the gray matter is located on the surface of the cerebral hemispheres as the cerebral cortex. A smaller portion of the gray matter is located in the depth of the cerebral, cerebellar, brainstem and spinal gray matter. Gray matter in the depth of the cerebral is organized into distinct nuclei, and we investigate segmentation of these nuclei in the thesis. We are in particular interested in a specific deep gray matter structure, the thalamus.

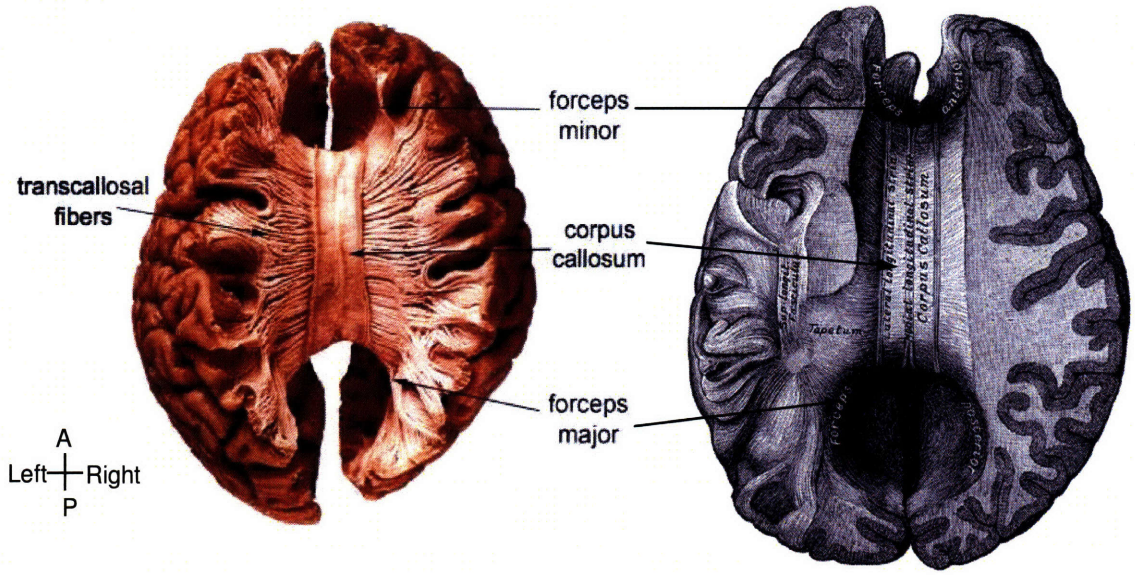

Figure 2-5: Corpus callosum from a dissected brain (left) and an illustration of the corpus callosum (right). The corpus callosum is a commissurral tract bundle that connects the two hemispheres of the brain. Images are adapted from [82, 25].

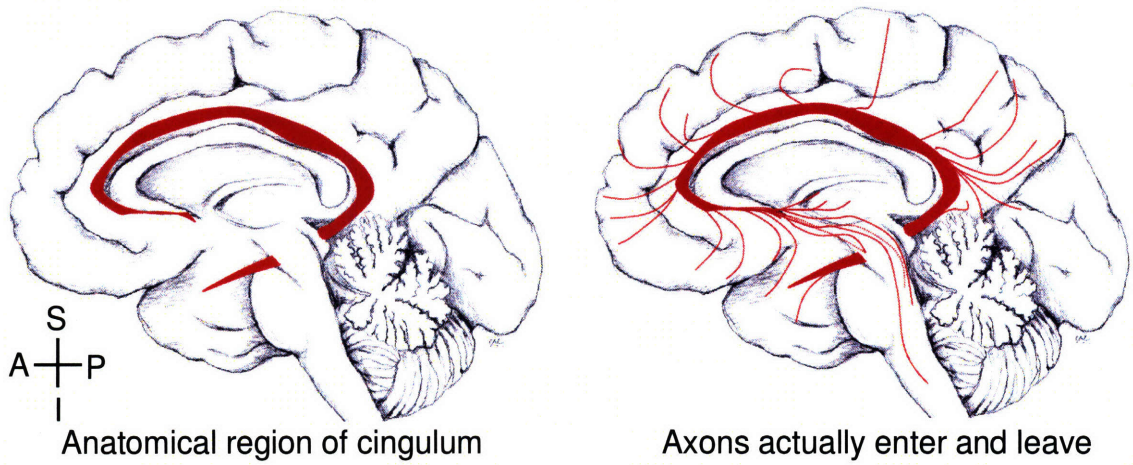

Figure 2-6: Illustration of the cingulum tracts. The left image illustrates the core of the cingulum, whereas the right image depicts axons entering and leaving this core. Images are adapted from [57].

Thalamus

A normal brain contains two thalami, one in each hemisphere, located symmetrically inferior to the corpus callosum and the fornix (Figure 2-8 (left)). Each thalamus is about 6 cm

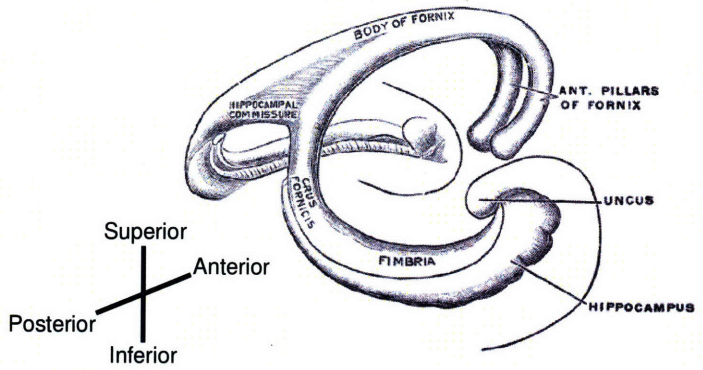

Figure 2-7: Illustration of the fornix. The fornix is part of the limbic system, connecting hippocampus to mammillary bodies. Image is adapted from [25].

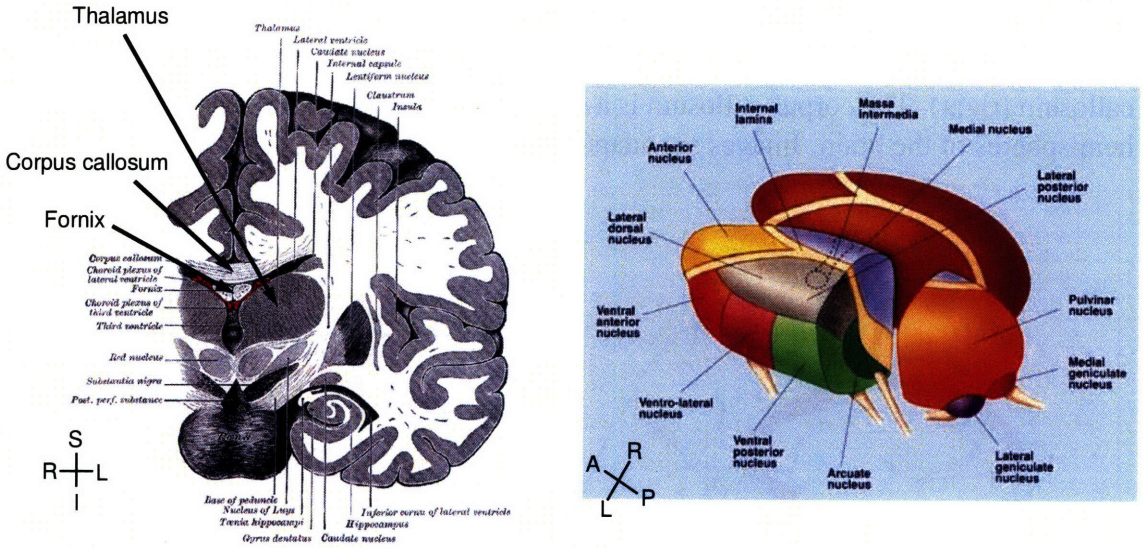

Figure 2-8: Illustration of the thalamus (left) and its distinct nuclei (right). The thalamus is the major relay center to the cortex for all sensations (except the sense of smell). Image is adapted from [25].

in length and exhibits a bean-like shape. All of the sensory pathways of the human brain (with the exception of the olfactory pathway, responsible for the sense of smell) project to the cortex via relay neurons in the thalamus. Hence, the thalamus acts as the central relay station of the brain. These relay neurons are clustered into discrete clusters (nuclei) that can be delineated based on histological or functional criteria (Figure 2-8 (right)).

2.4 Summary of the Chapter

In this chapter we briefly introduced the human brain, specifically the white matter bundles and the deep gray matter nuclei. Together, the structures that make up the human brain account for virtually all human activity, including movement, thought and emotions. However, the exact mechanisms of this process is not completely understood. Neuroscience aims to understand the human brain through studying the anatomy.

As we investigate in later chapters, a relatively new imaging technology, called diffusion weighted magnetic resonance imaging, provides adequate contrast for the automatic segmentation of the structures that are introduced in this chapter, and thus provide a new avenue for understanding the human brain anatomy.

Chapter 3

Diffusion Weighted Magnetic Resonance Imaging

In this chapter we provide a brief background on diffusion weighted magnetic resonance imaging (DW-MRI). We start by discussing free water diffusion, a phenomenon first observed by Robert Brown in 1828. We then relate this microscopic diffusion to macroscopic diffusion propagators that aims to quantify the Brownian motion in a probabilistic fashion.

Stejskal and Tanner were the first to measure water diffusion using the magnetic resonance technology in 1965. We briefly describe the experiment they performed and relate that to imaging water diffusion in the human brain.

This in-vivo imaging technique took 30 years to mature into the diffusion tensor imaging that we use today, and we briefly review the history of this development. And finally we explain how the diffusion tensor images are visualized and used to reconstruct the underlying white matter fiber tracts. We recommend [71, 12] to the interested reader for further understanding of the DW-MRI.

3.1 Diffusion

At the microscopic level, water molecules are constantly moving in a random motion, called Brownian motion. This motion is named after Robert Brown, who was the first to write about this phenomenon based on his observations on the movement of pollen grains in otherwise stationary water [11]. This microscopic level activity is also observed in a macroscopic level as diffusion. For example, a single drop of dye colors a whole glass of water in time. In that case, even though the dye molecules are moving in random directions at the microscopic level, this random motion translates into a macroscopic movement from higher dye concentration areas to the lower ones. Even without the dye, water molecules are diffusing among themselves, in a phenomenon called self-diffusion. This self-diffusion follows the same laws of diffusion and we use the word diffusion in the thesis to describe the self-diffusion of water molecules.

At the macroscopic level, diffusion is characterized with a diffusion propagator. The diffusion propagator is a probabilistic description of the distribution of molecule displacements in an ensemble. The probability of finding a molecule at the location described by vector $\mathbf{r}$, given that it was at location $\mathbf{r_o}$ at time τ ago is explained by the diffusion propagator, $p(\mathbf{r}|\mathbf{r_o}, \tau)$.

We also describe the diffusion within a volume v, with an ensemble average diffusion propagator, $p(\mathbf{r} - \mathbf{r_o}|\tau)$, where $\mathbf{r} - \mathbf{r_o}$ is the displacement vector. This ensemble average represents the average probability of a molecule, within the volume v, moving in the direction $\mathbf{r} - \mathbf{r_o}$, for a distance of $|\mathbf{r} - \mathbf{r_o}|$ in time τ. The DW-MRI aims to recover this ensemble average diffusion propagator for each voxel in the tissue being imaged.

3.2 Quantifying Diffusion: Pulse Gradient Spin Echo

In 1965 Stejskal and Tanner [69] were the first people to use magnetic resonance imaging to perform quantitative measurements of the molecular diffusion in a sample and relate this measurement directly to the diffusion coefficient. Detailed derivations of this relationship are presented in [69] and [12], but we review here the relevant findings that are necessary

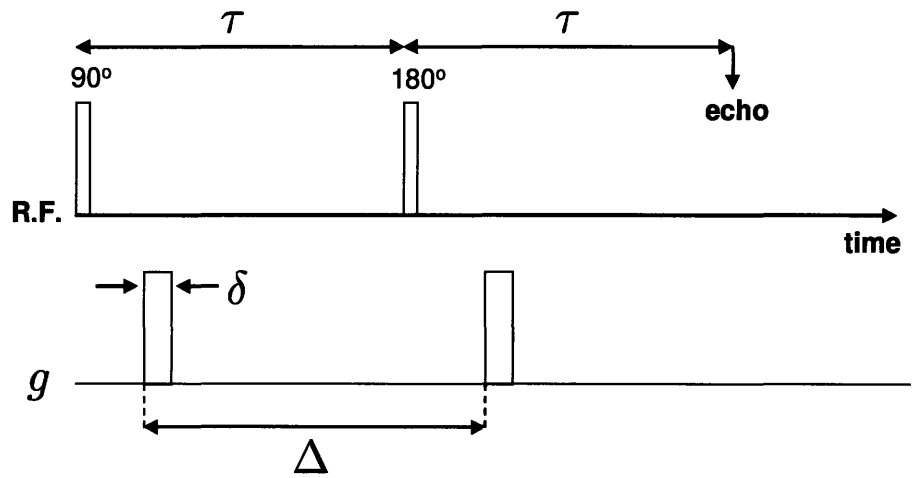

Figure 3-1: Pulse sequence diagram for the pulse gradient spin echo experiment. Please refer to the text for details.

to understand DW-MRI.

A pulsed gradient spin echo (PGSE) experiment works by encoding the location of an ensemble of molecules at an initial time, and then decoding the current locations of molecules after a short duration of time. If the molecules move during the experiment (for example due to diffusion), the encoding and decoding of the molecules at different locations result in a drop in the measured MRI signal.

The experiment starts by placing the sample under a strong magnetic field. Due to the strong magnetic field, the spins of the molecules within the sample get oriented along the magnetic field direction. First a radio frequency signal is applied to flip the spins to the transverse plane (Figure 3-1). This is followed by the first gradient pulse that causes a phase shift on the spins. This phase shift is not constant in space, but is varying according to a gradient direction and strength, thus encoding the current location of the molecules into the phase of the spins. A second radio frequency signal followed by a gradient pulse of the same direction, duration and strength of the first one is then applied. If there is no movement in the molecules along the gradient direction, the spins are expected to rephase exactly as before after this second gradient. However, in the case of molecular movement along the gradient direction, there is an attenuation in the MRI signal, quantifying molecular diffusion.

Stejskal and Tanner related this MRI signal attenuation to the diffusion coefficient. Assuming the molecules do not move during the application of the diffusion encoding gradient, the phase accumulated by each spin is:

$$\phi = \gamma\delta\mathbf{g}^T\mathbf{r}_o,$$

where γ is the Larmor constant, δ is the gradient duration, $\mathbf{g}$ is the gradient strength vector and $\mathbf{r}_o$ is the initial position of the molecule. The phase accumulated by each spin after the second gradient then becomes:

$$\phi = \gamma\delta\mathbf{g}^T(\mathbf{r} - \mathbf{r}_o),$$

where $\mathbf{r}$ is the position of the molecule when the second gradient is applied. An MRI signal measured at this experiment is an average magnetization of spins, each with phase ϕ_k:

$$S = S_o\left(\frac{1}{N}\sum_{k=1}^{N} e^{i\phi_k}\right), \tag{3.1}$$

where N is the number of spins in the ensemble, and S_o is the spin echo signal in the absence of any gradient directions. We rewrite Equation 3.1 in terms of a probability density function:

$$S = S_o\int e^{i\phi}p(\phi|\tau)d\phi,$$

where τ is the diffusion time. Let $\mathbf{R} \triangleq \mathbf{r} - \mathbf{r}_o$ and $\mathbf{q} \triangleq \gamma\delta\mathbf{g}$, then

$$\begin{aligned} S(\mathbf{q},\tau) &= S_o\int e^{i\mathbf{q}^T\mathbf{R}}p(\mathbf{q}^T\mathbf{R}|\tau)d\{\mathbf{q}^T\mathbf{R}\} \\ &= S_o\int e^{i\mathbf{q}^T\mathbf{R}}p(\mathbf{R}|\tau)d\mathbf{R} \\ &= S_oFT\{p(\mathbf{R}|\tau)\}, \end{aligned} \tag{3.2}$$

where $FT\{p(\mathbf{R}|\tau)\}$ is the Fourier transform of the probability density function. Therefore, the signal that is measured in the PGSE experiment directly relates to the Fourier transform of the probability density function of the spin displacements. This means we can vary

the diffusion encoding gradient direction, or equivalently $\mathbf{q}$, to obtain samples from this Fourier transform in order to construct the average diffusion propagator for the spins in the ensemble.

The PGSE experiment as described measures the ensemble average diffusion propagator for the whole sample. It is also possible to introduce varying magnetic gradients in the MRI machine in order to measure the signal attenuations for each voxel in the sample, in order to create 3D volumes [70, 42].

In the following sections we investigate how such images are created and processed in order to quantify water diffusion in the brain tissue.

3.3 Diffusion in Free Water

When there are no barriers to hinder the displacement of molecules, the Brownian motion of the molecules leads into isotropic diffusion. In this case, the amount of diffusion is related to the diffusion coefficient, which depends on the medium properties but not on direction. The diffusion coefficient was related to the mean square displacement by Einstein [20] with the following equation:

$$D = \frac{1}{6\tau} \left\langle \mathbf{R}^T \mathbf{R} \right\rangle ,$$

where τ is the diffusion time, and $\mathbf{R} = \mathbf{r} - \mathbf{r}_o$ is the net displacement vector. $\langle . \rangle$ denotes an ensemble average.

The isotropic diffusion is quantified by a Gaussian diffusion propagator with mean $\mathbf{r}_o$ and covariance matrix $2D\tau \times \mathbf{I}_3$, where $\mathbf{I}_3$ is the 3 dimensional identity matrix,

$$p(\mathbf{R}|\tau) = p(\mathbf{r} - \mathbf{r}_o|\tau) = \frac{1}{\sqrt{D(4\pi\tau)^3}} exp\left[\frac{-(\mathbf{r} - \mathbf{r}_o)D^{-1}(\mathbf{r} - \mathbf{r}_o)}{4\tau}\right].$$

When free diffusion is measured through the PGSE experiment of Section 3.2, the resulting MRI signal becomes:

$$\begin{aligned} S(\mathbf{q}, \tau) &= S_o FT\{p(\mathbf{R}|\tau)\} \\ &= S_o e^{-D\tau|\mathbf{q}|^2} \end{aligned} \tag{3.3}$$

Let $b \triangleq \tau|\mathbf{q}|^2$, which is called LeBihan's b-value, then 3.3 becomes

$$S(\mathbf{q},\tau) = S_o e^{-bD} \quad . \tag{3.4}$$

This results means the diffusion coefficient can be quantified through two measurements: the first one with no gradient encoding, called the $b = 0$ image, and another measurement with any gradient direction. The resulting diffusion coefficient is

$$D = -\frac{1}{b} ln \frac{S}{S_o}. \tag{3.5}$$

3.4 Diffusion in Tissue

Water diffusion is hindered in biological tissue due to barriers such as cell membranes or myelinated axon sheaths (Figure 3-2). When these barriers are oriented coherently, such as in the cerebral white matter, water diffusion becomes faster in the direction along the fibers and slower across the fibers [6].

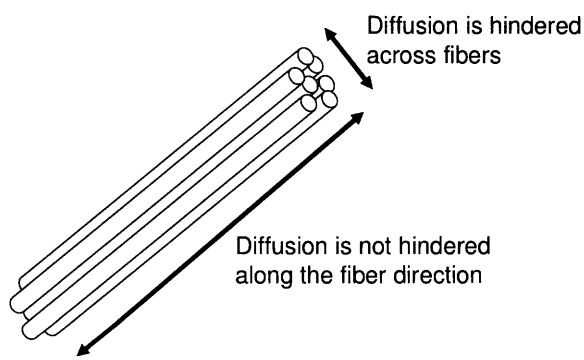

Figure 3-2: Illustration of anisotropic diffusion in fibrous tissue. Diffusion is hindered in biological tissue due to barriers such as cell membranes or myelinated axon sheaths.

In the case of biological tissue, the diffusion coefficient is no longer the same in every direction, and the resulting diffusion is called anisotropic. In the anisotropic diffusion case,

we can no longer calculate the diffusion coefficient with Equation 3.5, since the measured signal amplitude is no longer independent from the encoding gradient direction. The measured diffusion coefficient in this case is called the apparent diffusion coefficient (ADC), to distinguish the anisotropic diffusion in tissue from the isotropic diffusion of free water.

Measuring the apparent diffusion coefficient for each voxel in biological tissue leads into construction of ADC images [70, 42]. This development became significant when it was determined that the ADC is reduced by a significant amount within minutes of a stroke, whereas other MRI modalities remain constant for hours [54]. In fact, ADC images currently remain the most effective indicator of stroke in the emergency room setting [38].

Relaxing the assumption of isotropy, but restricting the diffusion propagator to a 3D Gaussian results in diffusion tensor imaging (DTI) [6]. It was Peter Basser who introduced DTI in 1994 [6], and DTI remains the most popular DW-MRI model that is currently in use. Further relaxing the Gaussian assumption allows for more accurate representations of the diffusion propagator, but due to technical limitations, these models are still an active research area.

3.5 Diffusion Tensor Imaging

Diffusion tensor imaging assumes that the diffusion propagator takes a Gaussian form, with zero mean and covariance matrix $2\mathbf{D}\tau$,

$$p(\mathbf{R}|\tau) = p(\mathbf{r} - \mathbf{r}_o|\tau) = \frac{1}{\sqrt{|\mathbf{D}|(4\pi\tau)^3}} exp\left[\frac{-(\mathbf{r} - \mathbf{r}_o)\mathbf{D}^{-1}(\mathbf{r} - \mathbf{r}_o)}{4\tau}\right],$$

where τ is the diffusion time, $\mathbf{r}$ is the final position of the diffusing water molecule and $\mathbf{r}_o$ is the original position of the molecule. $\mathbf{D}$ is the positive definite apparent self diffusion tensor, which is in general referred as the diffusion tensor. Given the diffusion tensor, the apparent diffusion coefficient $D_{\mathbf{R}}$ for any given directional vector $\mathbf{R}$ is calculated as:

$$D_{\mathbf{R}} = \mathbf{R}^T\mathbf{D}\mathbf{R}.$$

3.5.1 Diffusion Tensor Estimation

Similar to the diffusion coefficient estimation of Section 3.3, we use the Fourier relationship between the Gaussian model and the PGSE signal, which in this case equals:

$$S(\mathbf{q}, \tau) = S_o e^{-\tau \mathbf{q}^T \mathbf{D} \mathbf{q}}. \tag{3.6}$$

Introducing unit vectors, $\mathbf{u} \triangleq \mathbf{q}/|\mathbf{q}|$, and b-value, $b = \tau|\mathbf{q}|^2$, we rewrite Equation 3.6 as

$$S(b, \mathbf{u}) = S_o e^{-b\mathbf{u}^T D \mathbf{u}}. \tag{3.7}$$

Finally, taking the log of both sides reduces Equation 3.7 into a set of linear equations:

$$\ln S(b, \mathbf{u}) = \ln S_o - \sum_{i=1}^{3} \sum_{j=1}^{3} \mathbf{b}_{ij} \mathbf{D}_{ij}, \tag{3.8}$$

where $\mathbf{b} = \{\mathbf{b}_{ij}\} \triangleq b\mathbf{u}\mathbf{u}^T$ is called the **b**-matrix. Equation 3.8 is a linear system of equations with 7 unknowns, which include S_o and 6 independent elements of the diffusion tensor $\mathbf{D}$ (since $\mathbf{D}$ is constrained to be symmetric). This system can be solved by measuring diffusion once with $b = 0$ for S_o, and 6 more times with a higher b-value (around $1000 smm^{-2}$) in 6 different gradient directions. However, since the MRI signals tend to be noisy it is a common practice to collect many more than 7 images, and then solve for the diffusion tensor through a linear least squares [34] or some other non-linear optimization method, while constraining the diffusion tensors to be symmetric and positive definite [48].

3.5.2 Eigensystem of Diffusion Tensor

The diffusion tensor's eigensystem is useful for visualization and quantitative analysis. For a given diffusion tensor $\mathbf{D}$, the corresponding eigenvectors and eigenvalues satisfy:

$$\mathbf{D} = \mathbf{E}\mathbf{V}\mathbf{E}^T = [\mathbf{e}_1|\mathbf{e}_2|\mathbf{e}_3] \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} [\mathbf{e}_1|\mathbf{e}_2|\mathbf{e}_3]^T,$$

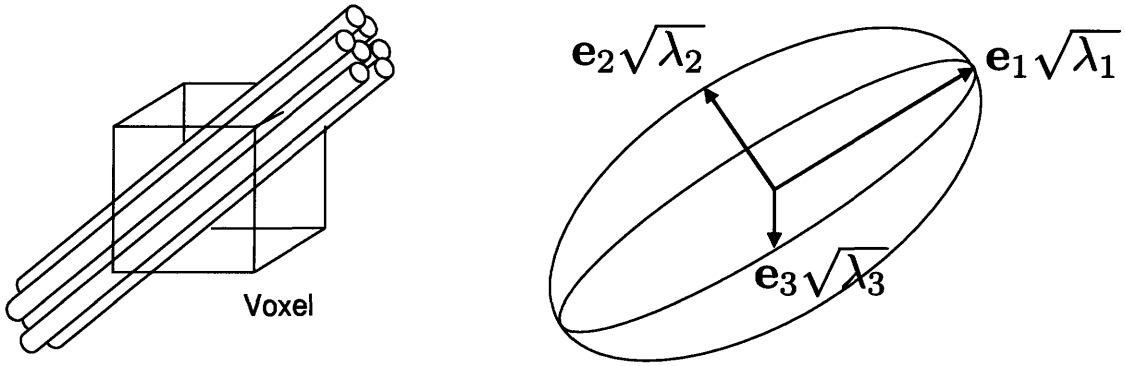

Figure 3-3: Illustration of diffusion tensor in fibrous tissue through an isoprobability surface. Since the underlying model is Gaussian, the isoprobability surface is an ellipse. The major axes of the ellipse are aligned with the eigenvectors of the diffusion tensor, and the radii are proportional to the $\mathbf{e}_i\sqrt{\lambda_i}$.

where $\mathbf{E}$ is a 3×3 matrix whose columns are the eigenvectors of the diffusion tensor, and $\mathbf{V}$ is a diagonal 3×3 matrix whose diagonal elements are the corresponding eigenvalues. $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$ are the eigenvalues of the diffusion tensor in decreasing amplitude and $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$ are the corresponding eigenvectors. The eigenvector with the largest eigenvalue, $\mathbf{e}_1$, is called the principal eigenvector and defines the major fiber tract axis of the tissue [6].

The three dimensional Gaussian diffusion propagator can be visualized through an isoprobability surface. Since the underlying model is Gaussian, the shape of the isoprobability surface is independent of the probability level chosen for this visualization, and the surface is described by an ellipse. The major axes of the ellipse are aligned with the eigenvectors of the diffusion tensor, and the radii are proportional to the $\mathbf{e}_i\sqrt{\lambda_i}$. Refer to Figure 3-3 for an illustration.

3.5.3 Scalar Derivatives of Diffusion Tensor

Unlike the ADC images, the diffusion tensor images are directional and matrix-valued. Therefore, the DTI can not be visualized and processed in the same way as gray-scale (scalar) structural MR images. This problem led to the introduction of scalar invariants, which are visualized and processed in a similar fashion to the gray scale images. There are two major types of scalar diffusion invariants, diffusivity measures and anisotropy mea-

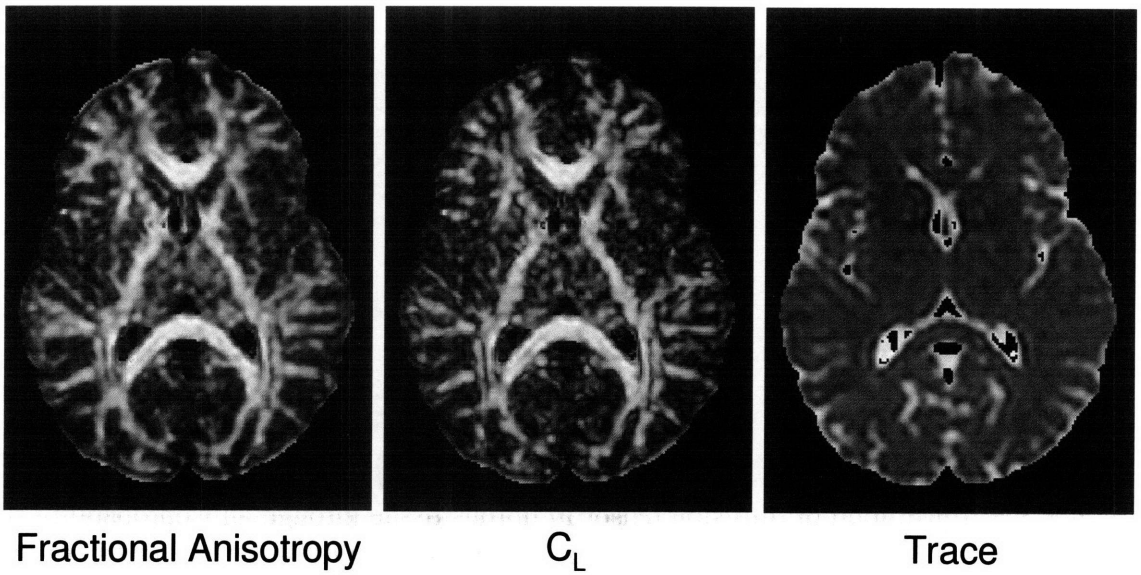

Figure 3-4: Common scalar derivatives of the diffusion tensor in an exemplar brain. White matter tracts become apparent in Fractional Anisotropy and C_L images as high intensity (white) areas, whereas the trace is nearly constant in most of the brain tissue.

sures:

Diffusivity Measures

Diffusivity measures are used to quantify the overall diffusion in a given voxel. The most common diffusivity measures relate to the trace of the diffusion tensor [4, 15, 41]:

$$trace(\mathbf{D}) = \sum_{i=1}^{3} \lambda_i.$$

The trace of a tensor quantifies how free the water is to diffuse and is significantly higher in CSF as opposed to the brain tissue. It is relatively constant in the rest of the brain, which includes gray and white matter (Figure 3-4).

Anisotropy Measures

Anisotropy measures are used to quantify the degree of preference in diffusion in the principle eigenvector direction as opposed to other directions. The most commonly used measures include a linear shape measure (C_L) [80, 79] and fractional anisotropy (FA) [8]. C_L

is defined as:

$$C_L = \frac{\lambda_1 - \lambda_2}{\lambda_1}.$$

If the diffusivity in the principal diffusion direction, which is quantified by λ_1, is larger than diffusivity in the secondary direction, λ_2, then the resulting C_L value becomes higher, indicating a preference in the diffusion direction. Consequently, the value of C_L is higher in white matter than that in gray matter (Figure 3-4).

The most commonly used anisotropy measure is FA [8], which is defined as:

$$FA = \frac{1}{\sqrt{2}} \frac{\sqrt{\sum_{i=1}^{3}(\lambda_i - \overline{\lambda})}}{\sqrt{\sum_{i=1}^{3} \lambda_i^2}},$$

where $\overline{\lambda} = (1/3) \times (\lambda_1 + \lambda_2 + \lambda_3)$. Similar to the C_L, FA becomes higher with a relatively higher value of the principal eigenvalue λ_1, and is in general higher in the white matter (Figure 3-4).

3.6 Higher Order Diffusion Models

It is also possible to relax the Gaussian assumption on the diffusion propagator to describe non-Gaussian diffusion distributions. There are several methods designed for this purpose. All these methods sample the q-space, the Fourier space spanned by the gradient directions, in many directions and solve the inverse Fourier transform through either parametric or non-parametric solvers. These methods include Q-space imaging [12], Diffusion Spectrum Imaging [76] and Q-Ball imaging [71].

3.7 Tractography

With the introduction of diffusion tensor imaging, an interesting area of research has emerged, concerned with reconstruction of the underlying tracts from the diffusion tensor images. The common name for these algorithms is tractography. There are many methods in the literature that perform this task, but the central theme of these methods is to construct 3D

curves that follow the principal diffusion direction in the diffusion tensor data.

Streamline Tractography

The most common approach to tractography is the streamline method [7, 16, 53, 78]. These methods create 3D curves by starting at an initial location, called a seed voxel. From this initial location, the streamline methods extend the curve at small steps, typically smaller than the voxel dimension, following the principal diffusion direction at each step. Since the diffusion tensor data is discrete, the algorithms use an interpolation method to calculate an estimated tensor for any given location.

3.8 Summary of the Chapter

In this chapter, we presented a brief background on diffusion MR imaging. Water diffusion is a microscopic phenomenon that can be explained by diffusion propagators in a macroscopic scale. When measured using MR technology, this water diffusion results in an MR signal that directly relates to the Fourier Transform of the diffusion propagator.

Assuming different models for the diffusion propagators leads to different kinds of images. An isotropic diffusion assumption leads to apparent diffusion coefficient images, which are used as early indicators of stroke in the emergency room setting. A Gaussian propagator is the most commonly used for anisotropic diffusion, and results in diffusion tensor images. Diffusion tensor images are matrix-valued images unlike the structural MR images and therefore require development of new techniques for visualization and processing, which is the subject of the thesis.

Chapter 4

Unsupervised Clustering as a Deep Gray Matter Segmentation Tool

Recent work has shown that diffusion tensor imaging (DTI) can help resolve thalamic nuclei based on the characteristic fiber orientation of the corticothalamic/thalamocortical striations within each nucleus. In this chapter we describe a novel segmentation method based on spectral clustering. We compare several tensor similarity metrics, and we explicitly minimize the normalized cut criteria of the spectral clustering for a better optimization. Using this modified spectral clustering algorithm, we can resolve the organization of the thalamic nuclei into groups and subgroups based solely on the voxel affinity matrix, avoiding the need for explicitly defined cluster centers. Identifying nuclear subdivisions facilitates localization of functional activation and pathology to individual nuclear subgroups.

4.1 Introduction

All of the sensory pathways of the human brain (with the exception of the olfactory pathway) project to the cortex via relay neurons in the thalamus. Hence, the thalamus acts as the central relay station of the brain. These relay neurons are clustered into discrete groups (nuclei) that can be delineated based on histological or functional criteria.

Precise identification of the thalamic nuclei is essential in a clinical setting, since many motor-control disorders are surgically corrected by applying chronic electrical stimulation to the appropriate functional area of the thalamus. Currently, these regions are detected qualitatively before the operation using generic atlases along with structural MRI [26], which does not provide adequate contrast to identify distinct nuclei. Changes have also been reported in the thalamic nuclei during the progression of a large number of diseases, including schizophrenia [63] and Parkinson's disease [23].

Diffusion tensor imaging (DTI) is a relatively new imaging modality that measures free water diffusion, i.e. Brownian motion, of the endogenous water in tissue [6]. In human brain tissue, the water diffusion is not the same in all directions, since it is obstructed by structural elements such as cell membranes or myelin [6]. When this obstruction constrains the water diffusion in a coherent direction, such as within the cerebral white matter, the resulting water diffusion tensor becomes anisotropic, containing information about the directionality of the underlying tissue.

Since functionally related pathways target the same region of cortex once they leave the thalamus, they result in organization of diffusivity within the thalamus. This organized diffusion can be measured in DTI, and it has been proposed that the thalamic nuclei can be distinguished by their characteristic diffusion orientation [81]. On DTI, the thalamus shows distinct clusters of diffusion orientation (Figure 4-1). These clusters correspond to the anatomic location and fiber orientation to the thalamic nuclei [18] (Figure 4-2). Diffusion orientation thus provides a new anatomic criterion for distinguishing thalamic nuclei.

Although DTI can resolve thalamic nuclei, a segmentation algorithm is required to explicitly delineate the nuclei from the DTI data. Several methods have been proposed to resolve the thalamic nuclei, including the use of k-means [81], level-sets [33] and tract

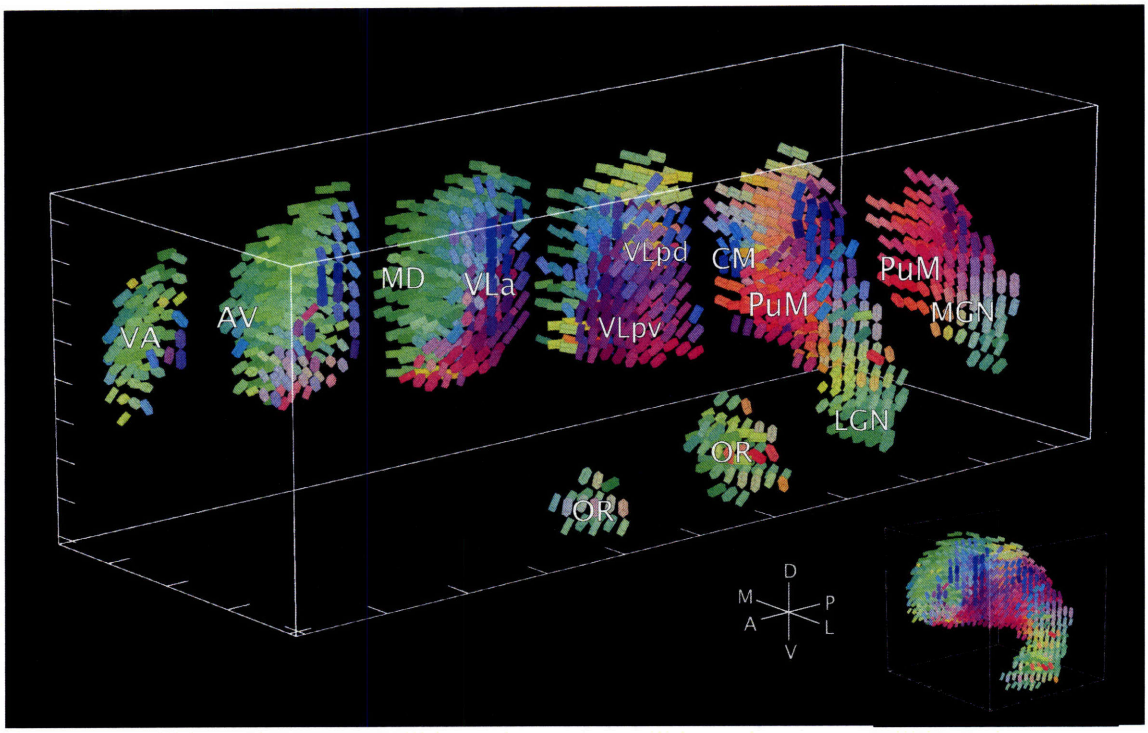

Figure 4-1: Three-dimensional DTI rendering of a thalamus. The cuboids depict the full diffusion tensor. The axes of the cuboids are scaled according to the tensor eigenvalues and are oriented according to the direction of the corresponding eigenvectors. The color indicates the direction of the principal eigenvector with red indicating medial-lateral, green anterior-posterior, and blue superior-inferior. Anatomic labels for the nuclei are shown on the thalamic hemisphere.

reconstructions from manually defined regions on the cortical surface [9]. A weakness of the k-means algorithm is its geometric bias towards ellipsoidal clusters. Further, both k-means and level sets show sensitivity to initialization, and susceptibility to local minima.

In this Chapter, we describe a new approach for segmentation of thalamic nuclei that builds on the spectral segmentation algorithm described by Shi and Malik [66] with some modifications. The spectral segmentation algorithm has attracted considerable interest in the pattern recognition community due its computational simplicity, strong empirical performance, and rich underlying theory [29, 37, 77]. The algorithm is based on a classical result from spectral graph partitioning which relates the second smallest eigenvector of the Laplacian matrix of a graph to optimal partitions. The spectral clustering algorithm has a number of desirable features. For example, the algorithm consists entirely of direct ma-

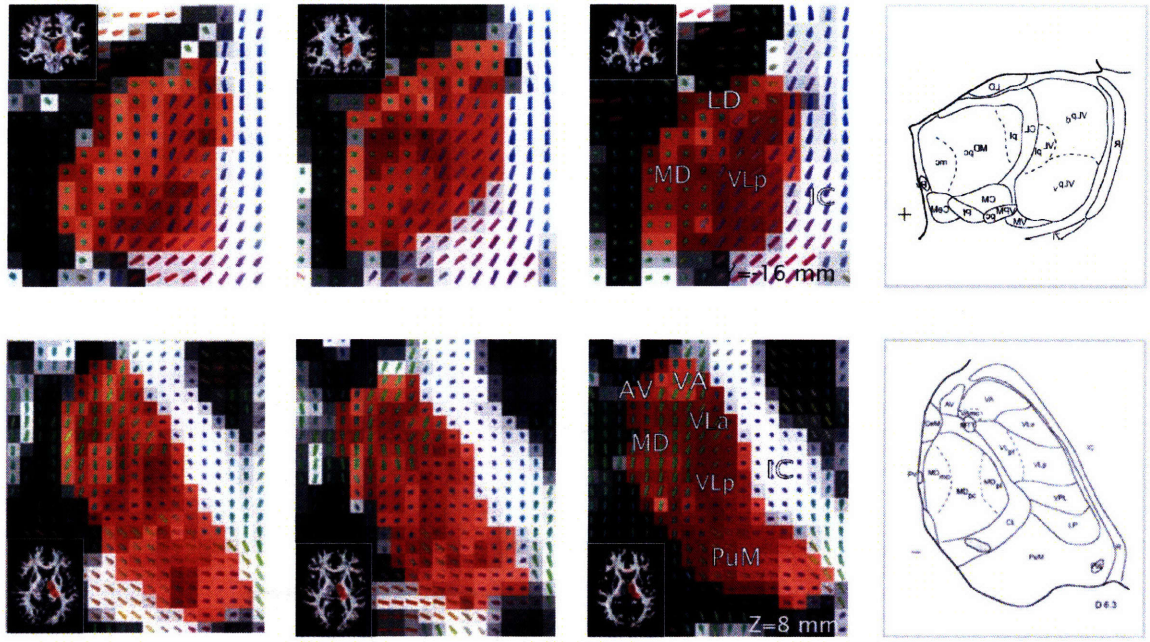

Figure 4-2: Coronal (top row) and axial (bottom row) DTI maps of 3 representative participants (columns). The rightmost column shows the corresponding camera lucida drawings from the Morel stereotactic atlas [52] (top: coronal section 9 mm anterior to PC level. bottom: axial section 6.3 mm dorsal to AC-PC plane). The DTI maps are normalized to MNI space. The axial section is from MNI Z=8 mm and the coronal section is from MNI Y = -16 mm. The cuboidal glyphs depict the diffusion tensor principal eigenvector within each voxel. The color scheme is the same as in Figure 4-1. The background grayscale image is the FA map. The red overlay on the FA map indicates the manually drawn mask for each individual. Anatomic labels for the nuclei are shown on the third participant's DTI maps.

trix operations and is therefore computationally efficient. Furthermore, the algorithm does not require a geometric representation of the clusters and therefore contains no explicit geometric biases.

In Section 4.2, we provide some background on spectral clustering and k-means algorithms. We then describe our modifications to the spectral clustering algorithm for DTI data. Finally, we present clustering results obtained on the DTI data from 10 healthy participants. We also provide results from a set of experiments to evaluate the performance of the spectral clustering algorithm. In these experiments, we measure the performance by quantifying the volume overlap between the clustering results and expert labels. We compared the performance of the spectral clustering method against an earlier method based on the k-means algorithm [81] as a benchmark.

4.2 Background

In this section we will provide a brief background on the k-means algorithm and spectral clustering. We will first review the k-means algorithm and provide the details of its application to DTI segmentation as proposed by Wiegell, et al. [81]. We will then review two spectral clustering methods that have different objective functions, but offer the same solution to two different problems. These methods share the same solution, because spectral clustering is suggested as a heuristic algorithm for NP-complete problems. This provides the basis for the modifications that we make to spectral clustering for use in DTI segmentation.

Some Notation

Before we begin, we would like to review the notation we use in this chapter. $N(\mathbf{v})$ represent the number of elements in vector $\mathbf{v}$ and $N(S)$ represent the number of elements in set S. A cluster c is represented by a set of indexes $I_c = \{i_1^c, i_2^c, \ldots, i_{N(S_c)}^c\}$, where $N(S_c)$ refers to the number of elements in cluster c. As follows, the overall dataset is represented by a union of clusters, $I = \{I_1, I_2, \ldots, I_m\}$. Similarly a cluster c contains a set of objects, which are represented by vertices in the graph notation of spectral clustering, $V_c = \{v_i | i \in I_c\}$.

4.2.1 k-means

The k-means algorithm is a clustering algorithm to separate n objects into k partitions, based on a distance measure defined for the objects [27, 10]. The algorithm starts with an initial set of k centroids, one for each cluster, and iteratively minimizes an objective function, also referred as the "distortion function." The iterations correspond to the estimation of a specific Gaussian mixture model using the expectation minimization algorithm. The corresponding Gaussian mixture model shares the same fixed covariance matrix for each of the clusters. The k-means algorithm is guaranteed to converge to a local optima in a finite number of iterations.

Inputs for the algorithm

The k-means algorithm requires the following four specifications:

1. k, the number of clusters,
2. a distance metric,
3. an initial set of cluster centroids,
4. a convergence criteria.

The number of clusters, k, and the distance metric are chosen by the user for the specific problem. The initial set of cluster centroids are usually chosen randomly. Since the algorithm is generally fast and guaranteed to converge to a local optima, the algorithm is usually run with a number of different random initializations, and the solution with the minimal distortion measure is chosen at the end. The convergence criteria is usually chosen to stop the algorithm when no significant gains are made to the distortion function. The pseudo-code for the algorithm is provided in Algorithm 4.1.

Algorithm 4.1 Pseudo-code for the k-means algorithm. $\mathbf{x}_i$ denotes input data.

```
Require: Number of clusters, k, and initial set of cluster centroids, {μ_1, μ_2, ..., μ_k}
  repeat
    for i = 1 to n do
      for c = 1 to k do
        Calculate distance between x_i and cluster centroid μ_c
      end for
      Assign x_i to the cluster with closest centroid
    end for
    for c = 1 to k do
      Adjust centroid μ_c to be the mean of objects in cluster c
    end for
  until Convergence criteria is met
```

Distortion Function

The distortion function that the k-means algorithm minimizes is a function of the cluster assignments of each of the objects being clustered:

$$D = \sum_{c=1}^{k} \sum_{\mathbf{x}_i \in \mathbf{S}_c} (\mathbf{x}_i - \mu_c)^2,$$

where there are k clusters, $\mathbf{S} = \{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_k\}$, and μ_c is the centroid for the cth cluster.

k-means as a segmentation tool in DTI

The k-means algorithm is used as a segmentation tool for deep gray matter in DTI by Wiegell, et al. [81]. For the segmentation, Wiegell, et al. propose a distance metric based on the Mahalanobis spatial distance and the Frobenius tensor distance. More specifically, the distance metric E_{ic}, between voxel i and cluster c, is set to be:

$$E_{ic} = \|\mathbf{x}_i - \overline{\mathbf{x}}_c\|_{\mathbf{W}_c} + \gamma \|\mathbf{D}_i - \overline{\mathbf{D}}_c\|,$$

where $\mathbf{x}_i$ is the location of voxel i, and $\overline{\mathbf{x}}_c$ is the mean voxel location for cluster c. Similarly $\mathbf{D}_i$ is the diffusion tensor of voxel i, and $\overline{\mathbf{D}}_c$ is the mean diffusion tensor for cluster c. $\mathbf{W}_c$ is the covariance matrix for the locations of voxels in cluster c, and γ is a weighting parameter that controls the tradeoff between the voxel location distance and tensor distance.

Since the k-means algorithm assumes the same isotropic covariance matrix for each of the clusters, the resulting clusters have a spherical bias. To avoid this problem, the authors proposed the Mahalanobis norm, which is defined as:

$$\|\mathbf{x}\|_{\mathbf{W}} = (\mathbf{x}^T \mathbf{W}^{-1} \mathbf{x})^{-1/2}.$$

The covariance matrices, $\mathbf{W}_c$, are calculated at every iteration. The distance between the tensors are calculated using the Frobenius norm, which is defined as:

$$\|\mathbf{F}\| = [\mathrm{Tr}(\mathbf{F}^T \mathbf{F})]^{1/2},$$

where $\mathrm{Tr}(.)$ represents the trace of a matrix. Please note that since the diffusion tensors are symmetric, this operation is equivalent to first vectorizing the tensors (converting the 3×3 matrix into a 9×1 vector) and then calculating the Euclidean distance between the two vectors. The vectorized tensor is given by

$$\mathbf{d} = [\mathbf{D}(1,1)\ \mathbf{D}(1,2)\ \mathbf{D}(1,3)\ \mathbf{D}(1,2)\ \mathbf{D}(2,2)\ \mathbf{D}(2,3)\ \mathbf{D}(1,3)\ \mathbf{D}(2,3)\ \mathbf{D}(3,3)]^T,$$

where the right hand side of the equation takes into account that the diffusion tensor $\mathbf{D}$ is symmetric. The weighting parameter γ is calculated to balance the covariance matrices of the voxel location and tensor distance metrics:

$$\gamma = \left(\frac{\mathrm{Tr}(\mathbf{\Sigma_x})}{\mathrm{Tr}(\mathbf{\Sigma_d})}\right)^{1/2},$$

where $\mathbf{\Sigma_x}$ is the covariance matrix for the voxel locations and $\mathbf{\Sigma_d}$ is the covariance matrix for the vectorized tensors.

A random initialization for the k-means algorithm is not desirable in this medical setting, since the resulting clustering depends heavily on the initialization, and obtaining different segmentations on the same dataset with the same algorithm degrades the confidence in the algorithm's results. Therefore, Wiegell, et al. propose a uniform initialization method, where the centroids are chosen to uniformly span the range of voxel locations. For this purpose, Wiegell, et al. uniformly sample the line that passes through the posterior tip of the thalamus and also the center of mass within the plane spanned by the first and second eigenvectors of the voxel location covariance matrix.

4.2.2 Spectral Clustering

Spectral clustering is a clustering algorithm to separate n objects into k partitions. The algorithm starts by converting the data into a weighted graph $G = (V, E)$, where every object is a vertex in V and the distances between objects are placed in a $n \times n$ edge-weight matrix $\mathbf{W}$. The algorithm only uses the pairwise distance measures in the edge-weight matrix $\mathbf{W}$, and therefore it can utilize arbitrary similarity measures as distances. Once a

graph is constructed, the algorithm calculates a k-way graph portion using the eigenvectors of the normalized weight matrix.

Normalized Cuts

Shi and Malik define a graph for application of spectral clustering to image segmentation by taking each pixel as a vertex of the graph [67]. For each node pair (i, j), the corresponding edge weight ω_{ij} is calculated from a feature similarity term $f(\mathbf{x}_i - \mathbf{x}_j)$ and a spatial proximity term $s(\mathbf{x}_i, \mathbf{x}_j)$:

$$\omega_{ij} = \exp\left(\frac{-f(\mathbf{x}_i - \mathbf{x}_j)}{\sigma_f}\right) \times \begin{cases} \exp\left(\frac{-s(\mathbf{x}_i, \mathbf{x}_j)}{\sigma_s}\right), & \text{if } s(\mathbf{x}_i, \mathbf{x}_j) \leq r \\ 0, & \text{otherwise,} \end{cases}$$

where an exponential kernel is used in the edge-weight calculation and σ_f and σ_s are parameters to control the trade-off between the feature and spatial distance measures. To reduce the size of graph from $n \times n$ to a more reasonable size for large problems, no edges are placed between objects (pixels) that are further than distance r away from each other.

Once a graph is constructed, one can calculate the k-way graph portion using the Normalized Cut (NCut) criteria. Formally, the problem is now reduced to the following:

Given a graph, $G = (V, E)$ and its corresponding edge-weight matrix $\mathbf{W} = \{\omega_{ij}\}$, find a set of disjoint sets $\mathbf{V}_1,\ \mathbf{V}_2, \ldots, \mathbf{V}_k$, such that $\bigcup_{i=1}^{k} \mathbf{V}_i = \mathbf{V}$, which minimizes the normalized cut:

$$\text{NCut}\,(\mathbf{V}_1,\ \mathbf{V}_2, \ldots,\ \mathbf{V}_k) = \sum_{i=1}^{k} \frac{\text{asso}(\mathbf{V}_i, \mathbf{V}) - \text{asso}(\mathbf{V}_i, \mathbf{V}_i)}{\text{asso}(\mathbf{V}_i, \mathbf{V})} \quad ,$$

where $\text{asso}(\mathbf{A}, \mathbf{B}) = \sum_{i \in \mathbf{A}, j \in \mathbf{B}} \omega_{ij}$ [66]. NCut is a preferred measure over a more traditional minimum cut, since the Ncut criteria results in more even sized segmentations as opposed to the minimum cut criteria [67].

Finding the NCut, even for the case where $k = 2$, is NP-complete due to the combinatorial nature of the discrete permutations. Therefore, we need a polynomial time algorithm that will provide a reasonable approximation. Shi and Malik describe a bi-partioning al-

gorithm that uses the second eigenvector of the affinity matrix and prescribe using it iteratively to accomplish the required k-way cut. Shi and Malik provided an elegant reasoning for their algorithm and it is summarized below for the convenience of the reader.

Let $\mathbf{x}$ be an indicator vector, such that $\mathbf{x}_i = 1$ if node i is in $\mathbf{A}$, and $\mathbf{x}_i = -1$ if node i is in $\mathbf{B}$, and let $d_i = \sum_j \omega_{ij}$ be the degree of node i. Also let $\mathbf{D}$ be a diagonal matrix with $\mathbf{D}(i,i) = d_i$. Then it follows that

$$\text{NCut}(\mathbf{A},\mathbf{B}) = \frac{\text{asso}(\mathbf{A},\mathbf{V}) - \text{asso}(\mathbf{A},\mathbf{A})}{\text{asso}(\mathbf{A},\mathbf{V})} + \frac{\text{asso}(\mathbf{B},\mathbf{V}) - \text{asso}(\mathbf{B},\mathbf{B})}{\text{asso}(\mathbf{B},\mathbf{V})} ,$$

which reduces to (after a considerable amount of algebra [67]):

$$\text{NCut}(\mathbf{A},\mathbf{B}) = \frac{\mathbf{y}^T(\mathbf{D}-\mathbf{W})\mathbf{y}}{\mathbf{y}^T\mathbf{D}\mathbf{y}} , \tag{4.1}$$

such that $\mathbf{y}_i \in \{1, -b\}$ and $\mathbf{y}\mathbf{D}\mathbf{1} = 0$, where $\mathbf{b} = \sum_{j\in\mathbf{A}} \mathbf{d}_j / \sum_{j\in\mathbf{B}} \mathbf{d}_j$ and $\mathbf{1}$ is a vector of all ones.

The expression in Equation 4.1 is the Rayleigh quotient and if $\mathbf{y}$ is relaxed to take on continuous values, it is minimized by solving the generalized eigenvalue system:

$$\begin{aligned}
(\mathbf{D}-\mathbf{W})\mathbf{y} &= \lambda\mathbf{D}\mathbf{y}, \\
\mathbf{D}\mathbf{y} - \mathbf{W}\mathbf{y} &= \lambda\mathbf{D}\mathbf{y}, \\
-\mathbf{W}\mathbf{y} &= (\lambda-1)\mathbf{D}\mathbf{y}, \\
\mathbf{D}^{-1}\mathbf{W}\mathbf{y} &= (1-\lambda)\mathbf{y}, \\
\mathbf{M}\mathbf{y} &= \zeta\mathbf{y},
\end{aligned}$$

where $\mathbf{M} = \mathbf{D}^{-1}\mathbf{W}$ and $\zeta = (1-\lambda)$. $\mathbf{M}$ is a Markovian probability matrix that describes the one-step probability of transitions, if the graph is thought of as a probabilistic graph, where each edge weight represents the probability of moving from one vertex to another. Therefore, the largest eigenvalue of the matrix $\mathbf{N}$ is 1 and the corresponding eigenvector is a constant vector. Furthermore, the other eigenvalues of the matrix are guaranteed to be less than 1 (if the graph is connected) and the corresponding eigenvectors are real and

orthogonal.

Thus, the Rayleigh quotient is minimized (non-trivially) by the second largest eigenvalue/eigenvector pair of $\mathbf{M}$, and the minimum value of NCut is $1 - \zeta$.

However, there is no guarantee that the scaled indicator vector $\mathbf{y}$ satisfies the first constraint $\mathbf{y}_i \in \{1, -b\}$. In fact in many real world problems, the continuous solution to $\mathbf{y}$ will not satisfy this constraint, and this point needs to be investigated further.

Distance Sensitive Ordering

Distance sensitive ordering (DSO) is another optimization problem that is based on the eigenvector ordering and solved through a spectral algorithm [18]. Even though the objective function is completely different from that of Shi and Malik, the solution is the same in the relaxed case (from a combinatorial problem to a continuous optimization problem). In DSO, the indicator vector consists of a stair function, unlike the step function of Shi and Malik. Their approach is summarized below for comparison with the Normalized Cut optimization.

Let $\mathbf{p}$ be a permutation vector (indicator vector) defined as $\mathbf{p}(1, 2, \ldots, n) = (p_1, p_2, \ldots, p_n)$, and for $\mathbf{W} = \{\omega_{i,j}\}$, the permuted affinity matrix is $\mathbf{W}_{per} = \{\omega_{p_i,p_j}\}$. An objective function $J(\mathbf{p})$ is defined as:

$$J(\mathbf{p}) = \sum_{l=1}^{n-1} \sum_{i=1}^{n-l} l^2 \omega_{p_i,p_j}.$$

With this objective function, the affinity matrix elements that are further away from the diagonal are heavily punished, thus forcing the affinity matrix to be ordered such that more similar voxels are located near each other. This problem is also potentially NP-Complete and thus we need an approximate solution.

To find a solution, let $\mathbf{q}$ be the shifted and scaled inverse index permutation vector:

$$q_i \propto \frac{p_i^{-1} - (n+1)/2}{n/2} \in \left\{ \frac{1-n}{n}, \frac{3-n}{n}, \ldots, \frac{n-1}{n} \right\},$$

which satisfies $\sum_i q_i = 0$ and $\sum_i q_i^2 = 1$. Thus, minimizing $J(\mathbf{p})$ is equivalent to mini-

mizing a second objective function $\overline{\mathbf{J}}(\mathbf{q})$, with constraints $\sum_i q_i = 0$ and $\sum_i q_i^2 = 1$.

$$\begin{aligned}
\overline{\mathbf{J}}(\mathbf{q}) &= \sum_{i,j} (q_i - q_j)^2 \omega_{p_i,p_j} \\
&= \sum_{i,j} (q_i^2 + q_j^2 - 2q_i q_j) \omega_{p_i,p_j} \\
&= 2\mathbf{q}^T(\mathbf{D} - \mathbf{W})\mathbf{q}.
\end{aligned}$$

However, it has been noticed experimentally [18] that the approximate solutions to the DSO problem greatly improve if the constraints on the weights are replaced with:

$$\begin{aligned}
\sum_i q_i d_i &= 0 \quad \text{and} \\
\sum_i q_i^2 d_i &= 1.
\end{aligned} \tag{4.2}$$

Minimization of $\overline{\mathbf{J}}(\mathbf{q})$ produces the following expression when the constraint of Equation 4.2 is applied using a Lagrangian multiplier:

$$\arg\min_{\mathbf{q}} \frac{2\mathbf{q}^T(\mathbf{D} - \mathbf{W})\mathbf{q}}{\mathbf{q}^T\mathbf{D}\mathbf{q}},$$

which is the Rayleigh quotient as it was the case with minimization of the NCut. If $\mathbf{q}$ is relaxed to take on continuous values, the solution is exactly the same as the NCut case.

4.3 Spectral Split & Merge Clustering

Normalized cuts and distance sensitive ordering aim to minimize significantly different objective functions; however, they prescribe the same solution to these two problems. Therefore, it is important to understand the subtle differences in the assumptions when the NP-complete problem of identifying the discrete indicator vectors is relaxed and solved by a continuous spectral approximation.

In the NCut case, the assumption is that the scaled indicator vector $\mathbf{y}$ contains values that are only at two discrete levels $\mathbf{y}_i \in \{1, -b\}$. In the distance sensitive ordering the scaled permutation vector $\mathbf{q}$, which is equivalent to $\mathbf{y}$, is assumed to contain a wider range of discrete values, $q_i \in C \times \left\{\frac{1-n}{n}, \frac{3-n}{n}, \ldots, \frac{n-1}{n}\right\}$, which represents a staircase function.

However, in a majority of the problems, the Fiedler vector becomes neither a step function nor a staircase function. Since the Fiedler vector is assumed to be a continuous approximation to the respective discrete indicator vector in each of the problems, it is necessary to further investigate the relationship between the approximate continuous solution and the discrete solution.

In Figure 4-3A, we present an ordered edge-weight matrix from a toy problem. The four clusters are clearly visible as four bright squares in the diagonals of the current permutation of the edge-weight matrix. The fourth cluster that is on the lower right side of the matrix is smaller in size than the other three and exhibits connectivity to the first and the third clusters but not to the second cluster.

The second eigenvector of the Markovian matrix is presented in Figure 4-3B. In this case, the second eigenvector is neither a perfect step function as assumed by the normalized cuts algorithm, nor a stair function as assumed by the distance sensitive ordering.

When the edge-weight matrix is permuted according to the second eigenvector, we obtain the matrix in Figure 4-3C. The two-way portioning according to the second eigenvector is outlined on the matrix with red squares. The resulting $NCut$ value is higher than the optimal value $NCut^*$. The optimal value $NCut^*$ is calculated in this simple case through an exhaustive search of all possible two-way portions. The difference in the $NCut$ value is due to the smaller fourth cluster being separated mistakenly with this approximate solution.

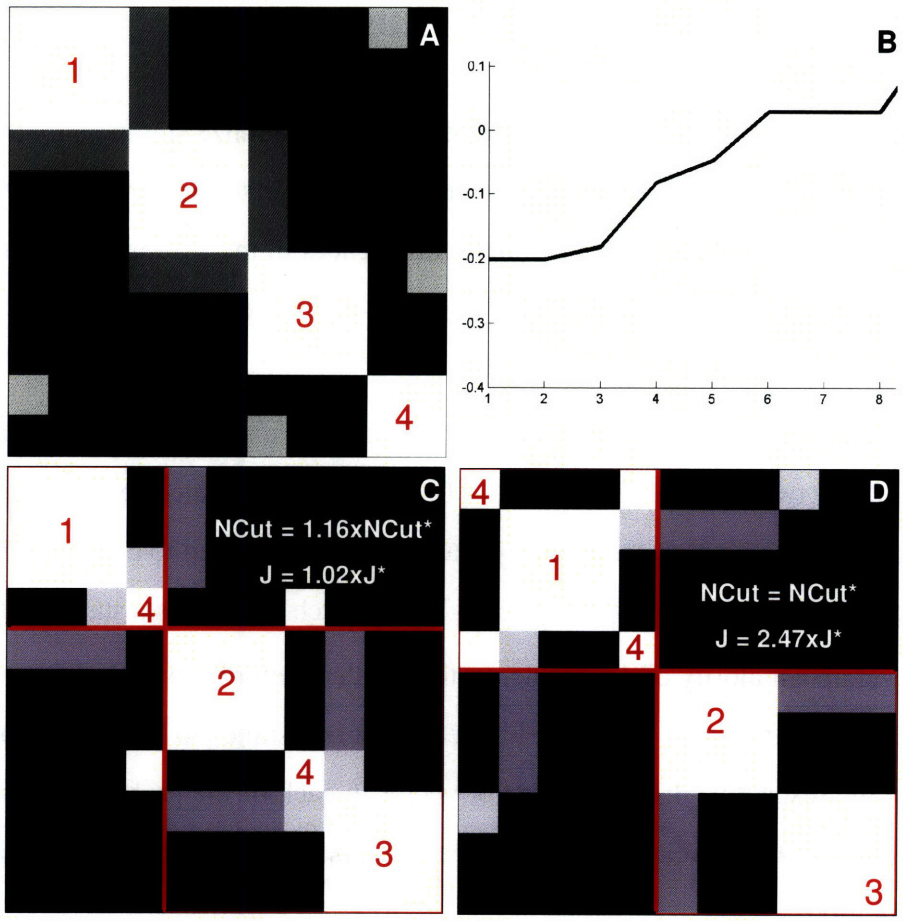

Figure 4-3: (A) An ordered edge-weight matrix from a toy problem. The four clusters are clearly visible as four bright squares in the diagonals of the current permutation of the edge-weight matrix. The fourth cluster that is on the lower right side of the matrix is smaller in size than the other three and exhibits connectivity to the first and the third clusters but not to the second cluster. The second eigenvector of the Markovian matrix is presented in (B). The second eigenvector resembles neither a step function nor staircase function. When the edge-weight matrix is permuted according to the second eigenvector, the fourth cluster is divided into two, which results in a non-optimal $NCut$ value as shown in (C). A greedy split and merge algorithm finds the optimal solution as shown in (D).

However, the DSO objective function J almost achieves the optimal value J^*. The optimal value J^* is once again calculated through an exhaustive search.

This example highlights the fact that the NCut and DSO are significantly different objective functions, and cannot be optimized by the same indicator vector as prescribed in [67] and [18]. Furthermore, as we presented, it is possible to construct simple toy examples where the second eigenvector offers a better solution for DSO, rather than for the NCut

problem. Due to the high dimensional nature of our problems, Shi and Malik's spectral clustering algorithm, when applied, often misassigns smaller clusters that exhibit high levels of connectivity to other clusters that are otherwise not very well connected as illustrated in Figure 4-3. To avoid this problem, we do not stop cutting when the number of clusters reaches k, the predetermined number of clusters. Rather, we continue cutting until a very high threshold of a 2-way $NCut$ value is reached. This threshold is typically close to 1, which is the level at which every point is clustered by itself regardless of the weights. This is referred as a spectral greedy splitting algorithm (GS) and is presented as pseudo-code in Algorithm 4.2. The greedy splitting is followed by a greedy merging algorithm, which is described in Algorithm 4.3, to reduce the number of clusters back to 1. The greedy merging algorithm produces a hierarchy tree of clusterings as it reduces the number of clusters back to 1. From the hierarchy tree, a k-way clustering is then extracted for the pre-determined number of clusters k. With the addition of the greedy merging algorithm, the misassignment of smaller clusters in high dimensional problems can be avoided as demonstrated in Figure 4-3C&D. We will further demonstrate the reduction achieved in $NCut$ through the new algorithm and its implications for segmentation accuracy later in Section 4.6 with DTI data.

At the final stage of our algorithm, we allow single object swaps between clusters to further reduce the NCut value. At every iteration of this stage, the algorithm considers the possibility of reassigning a single object to another cluster. If the resulting cut has a smaller $NCut$ value than before, that point is reassigned to its new cluster. The algorithm stops if there are no more objects left that can be reassigned to reduce $Ncut$. Pseudo-code for the object swaps is presented in Algorithm 4.4. The final iterations are guaranteed to converge within a finite number of stages since (1) the $NCut$ values are finite (bounded by m if there are m clusters) and (2) at every iteration the resulting graph's $NCut$ is reduced by at least the computer precision, which is also finite and non-decreasing. Since a point is reassigned only when the $NCut$ value is decreased, the resulting cut is guaranteed to have a $NCut$ value smaller than or equal to the original cut.

Algorithm 4.2 Pseudo-code for the spectral greedy splitting algorithm

Require: Graph $G = (V, E)$ and NCut threshold th
  Initialize: $I_1 = \{1, 2, \ldots, N(V)\}$, $I = \{I_1\}$ and $Istack \leftarrow \{I_1\}$ % Only one cluster
  while $Istack$ is not empty **do**
    $\widetilde{I} \leftarrow Istack(1)$ % Pick a subgraph to be divided
    $Istack \leftarrow Istack - Istack(1)$ % Remove the subgraph to be divided from $Istack$
    Calculate edge-weight matrix $\widetilde{\mathbf{W}}$ for $\widetilde{\mathbf{I}}$
    $\mathbf{M} \leftarrow \mathbf{D}^{-1}\widetilde{\mathbf{W}}$ % Calculate Markovian matrix
    Calculate the second largest eigenvector of $\mathbf{M}$, $\mathbf{v}_2$
    $[\mathbf{v}_{sorted}, \mathbf{ind}] = sort(\mathbf{v}_2)$ % Sort the second largest eigenvector
    $\widetilde{\mathbf{W}}_{sorted} = \widetilde{\mathbf{W}}(\mathbf{ind}, \mathbf{ind})$ % Permute $\widetilde{\mathbf{W}}$ according to the second largest eigenvector
    $\widetilde{I}_{sorted} = \widetilde{I}(\mathbf{ind})$ % Permute $\widetilde{I}$ according to the second largest eigenvector

    % Search for the best two way portioning of $\widetilde{\mathbf{I}}$
    $minNCut \leftarrow \infty$
    $bestcutpoint \leftarrow \infty$
    for $cutpoint = 1$ to $N(\widetilde{\mathbf{I}}) - 1$ **do**
      $newNCut \leftarrow NCutVal(G, \{\widetilde{I}_{sorted}(1 : cutpoint), \widetilde{I}_{sorted}(cutpoint + 1 : N(\widetilde{I})\})$
      if $newNCut < minNCut$ **and** $newNCut < th$ **then**
        $bestcutpoint \leftarrow cutpoint$
        $minNCut \leftarrow newNCut$
      end if
    end for

    if $bestcutpoint \neq \infty$ **then**
      $\widetilde{I}_1 = \widetilde{I}_{sorted}(1 : bestcutpoint)$
      $\widetilde{I}_2 = \widetilde{I}_{sorted}(bestcutpoint + 1 : N(\widetilde{I}))$
      $Istack \leftarrow \{Istack, \widetilde{I}_1, \widetilde{I}_2\}$ % Update $Istack$
      $I \leftarrow I - \widetilde{I}$ % Remove $\widetilde{I}$ since it is being cut into two clusters
      $I \leftarrow \{I, \widetilde{I}_1, \widetilde{I}_2\}$
    end if
  end while
  return Clustering results, I

Algorithm 4.3 Pseudo-code for the greedy merging algorithm

```
Require: Graph G = (V, E) and clustering results I = {I_1, I_2, ..., I_m}
  Initialize: Iind ← {1, 2, ..., m} and HieTree ← {}
  for iter = m to 2 × m − 1 do
    minNCut ← ∞
    for i = 1 to N(Iind) do
      for j = 1 to i − 1 do
        si ← Iind(i)
        sj ← Iind(j)

        % Consider putting together I_si and I_sj
        newNCut = NCutVal(G, {I − {I_si, I_sj}, {I_si, I_sj}})
        if newNCut < minNCut then
          minNCut ← newNCut
          besti = si
          bestj = sj
        end if
      end for
    end for
    % Update I by joining I_besti and I_bestj
    I ← I − {I_besti, I_bestj}
    I_iter+1 ← {I_besti, I_bestj}
    Iind = Iind − {besti, bestj} + {iter + 1}
    HieTree(iter − m + 1) ← {I_iter+1 ← {I_besti, I_bestj}}
  end for
  return HieTree
```

Algorithm 4.4 Pseudo-code for the object swaps algorithm

```
Require: Graph G = (V, E) and clustering results I = {I_1, I_2, ..., I_k}
  minNCut ← NCutVal(G, I)
  repeat
    for i = 1 to k do
      for j = 1 to N(I_i) do
        for all l such that l ∈ {1, 2, ..., k} − i do
          % Consider changing cluster assignment of object j from I_i to I_l
          newNCut ← NCutVal(G, {I − I_i − I_l, I_i − I_i(j), {I_l, I_i(j)}})
          if newNCut < minNCut then
            % Change cluster assignment of object j to cluster l
            I_i ← I_i − I_i(j)
            I_l ← {I_l, I_i(j)}
            minNCut ← newNCut
          end if
        end for
      end for
    end for
  until minNCut does not reduce in the last iteration
  return I
```

4.4 Application of Spectral Split & Merge Clustering to DTI

We treat the thalamic nuclei segmentation as a graph partitioning problem. For this purpose, we define a graph $G = (V, E)$, where vertices of the graph are chosen to be voxels from the DTI data and the corresponding edge weights are pairwise similarities of these voxels. The similarities are placed in an edge-weight matrix $\mathbf{W}$, where ω_{ij} is the weight of the edge between vertices V_i and V_j. Since the quality of the segmentation depends highly on the graph variables, they must be defined carefully.

4.4.1 Graph Construction

Our approach to nuclei segmentation is to partition the DTI data into compact regions with homogeneous diffusion properties, since similar diffusion properties are likely to be indicative of belonging to the same nuclei. For that purpose we are looking to construct a graph in which the vertices within a homogeneous region are well connected and the vertices in different homogeneous regions are not. This graph is constructed based on the spatial distances between the voxels as well as the tensor dissimilarities.

The first step in the construction is to choose an appropriate metric to quantify the tensor dissimilarities. There are a number of possible metrics for this purpose, such as the Frobenius norm that was proposed for the k-means clustering:

$$f_{\text{Frob}}\left(\mathbf{T}_i, \mathbf{T}_j\right) = \sqrt{\text{Tr}((\mathbf{T}_i - \mathbf{T}_j)^2)} \ .$$

$f_{\text{Frob}}\left(\mathbf{T}_i, \mathbf{T}_j\right)$ is explored in several DTI studies under names such as generalized tensor dot product [35], Frobenius norm [81], and Euclidean distance metric [1]. Another recently proposed metric is an information theoretic measure, namely the symmetric K-L divergence [72]:

$$f_{\text{KL}}\left(\mathbf{T}_i, \mathbf{T}_j\right) = \sqrt{\text{trace}(\mathbf{T}_i^{-1}\mathbf{T}_j + \mathbf{T}_j^{-1}\mathbf{T}_i) - 2n} \ ,$$

where $n = 3$ for diffusion tensors. Finally, we will use another metric that is based on the

angular distance between the principle eigenvector directions:

$$f_{\text{ang}}(\mathbf{T}_i, \mathbf{T}_j) = \arccos(|\mathbf{v}_i \cdot \mathbf{v}_j|) \quad ,$$

where $\mathbf{v}_i$ is the principal eigenvector of tensor $\mathbf{T}_i$. The absolute value of the dot product solves the problem with the sign ambiguity of the eigenvectors (if $\mathbf{v}_1$ is an eigenvector of $\mathbf{T}_i$, so is $-\mathbf{v}_1$). This angular metric was previously used to create scalar images from DTI and is called the intervoxel coherence [60]. Intervoxel coherence images are useful in quantifying the coherence of diffusion between neighboring voxels, and are generated using a sliding kernel and averaging angular distances between the center voxel and all neighboring voxels the kernel covers.

After choosing a tensor dissimilarity metric, we also need to incorporate spatial relations of the voxels into the graph. One way to do so is to linearly combine the tensor dissimilarity metric, $f(\mathbf{T}_i, \mathbf{T}_j)$, with a spatial distance, $s(\mathbf{x}_i, \mathbf{x}_j)$, as in [81]:

$$d((\mathbf{x}_i, \mathbf{T}_i), (\mathbf{x}_j, \mathbf{T}_j)) = f(\mathbf{T}_i, \mathbf{T}_j) + \gamma s(\mathbf{x}_i, \mathbf{x}_j) \quad ,$$

where γ is a weighting factor to control the trade-off between the tensor dissimilarity and the spatial distances. We choose γ in the same way as in [81]:

$$\gamma = \left(\frac{\text{Tr}(\mathbf{\Sigma_x})}{\text{Tr}(\mathbf{\Sigma_d})}\right)^{1/2} ,$$

where $\mathbf{\Sigma_x}$ is the covariance matrix for the voxel locations and $\mathbf{\Sigma_d}$ is the covariance matrix for the vectorized tensors.

Finally, the distance measures are converted into edge-weights through an exponential kernel:

$$\omega_{ij} = \exp\left(\frac{-d(\mathbf{x}_i - \mathbf{x}_j)}{\sigma}\right),$$

where σ is chosen to be the mean distance to the l-th neighbor of all vertices in the graph, and l was set to be 10 in all of the experiments described in this Chapter.

4.4.2 Graph Cuts

Once the graph is constructed, one can calculate the k-way graph partition using Normalized Cuts criteria. We solve this problem as described in Section 4.3.

4.5 Data Collection and Pre-Processing

4.5.1 Image acquisition

MRI data were acquired on ten healthy young participants (8 male, 2 female, age = 45 $\pm$ 10) at the Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital (Charlestown, Massachusetts). The data were acquired on a 3 Tesla Siemens Trio MRI scanner using an 8-channel head coil. All participants provided informed written consent by the guidelines of the MGH Internal Review Board. The participants gave their informed consent in writing prior to the session.

The DTI data were acquired using a twice-refocused spin-echo EPI sequence [64]. The sequence parameters were TR/TE=8400/82 ms, b=700 s/mm^2, gmax=26 mT/m, 10 T2 images, 60 diffusion gradient directions, 1 average, with total acquisition time 9 min 59 seconds. The 60 diffusion gradient directions were obtained from the electrostatic shell repulsion method [34]. Sixty four axial oblique slices were acquired. The slices were oriented in the intercommissural (AC-PC zero) plane. The field-of-view was 256$\times$256 mm and the matrix size was 128$\times$128 to give 2$\times$2 mm in-plane resolution. The slice thickness was 2 mm with 0 mm gap. The pre-averaged SNR of the T2 and diffusion-weighted images in deep white matter was, respectively, $\sim$ 28 and $\sim$ 18. The eddy current distortions between diffusion weightings were typically less than $\sim 1-2$ voxels.

4.5.2 Pre-processing

Correction for motion and residual eddy current distortion was applied by registering all of the scans to the first acquired non-diffusion-weighted scan for each participant. The registration was performed using the FLIRT (FMRIB's linear image registration tool) program which is available from the FSL library (http://www.fmrib.ox.ac.uk/fsl/flirt). The registration used a 12 degree-of-freedom global affine transformation and a mutual information cost function [30]. Trilinear interpolation was used for the resampling. The diffusion tensor, the tensor eigensystem, and the FA metric were calculated for each voxel using the formulas described in [5] and [61].

The diffusion tensor and FA volumes were normalized to MNI-space (Montreal Neurological Institute). The MNI-space normalization was preformed by registering each participant's T2 volume to a skull-stripped version of the MNI 152-subject T2 template [50] and then applying the transformation to the diffusion tensor and FA volumes. The MNI template was skull-stripped using the BET (brain extraction tool) program [68] from FSL (http://www.fmrib.ox.ac.uk/fsl/bet). The registration was performed using FLIRT with a 12 degree-of-freedom global affine transformation and a mutual information cost function [30].

The registration transformation was then applied separately the FA and tensor volumes. The FA volume was resampled using trilinear interpolation and the tensor volume was resampled using nearest neighbor interpolation. Nearest neighbor interpolation was used to resample the tensor volume to avoid partial volume averaging between different diffusion tensor orientations in neighboring voxels. The tensors were reoriented using the rotational portion $\mathbf{R}$ of the atlas transformation. The tensor reorientation was applied by pre- and post-multiplying the native-space diffusion tensors, i.e., $\mathbf{D} = \mathbf{R}\mathbf{D}\mathbf{R}^T$. The rotational portion of the atlas transformation $\mathbf{F}$ was extracted using canonical orthogonalization $\mathbf{R} = (\mathbf{F}^T\mathbf{F})^{-1/2}\mathbf{F}^T$ [2].

4.5.3 Thalamus Mask

Thalamus masks were drawn manually for each individual by a trained neuroanatomist [1]. The masks were drawn for each hemisphere on each individual's MNI-normalized FA map. The masks were drawn using TkMedit, which is part of the Freesurfer toolkit (http://surfer.nmr.mgh.harvard.edu). The thalamus was demarcated on axial, coronal and sagittal views of the FA map for each participant. Brightness and contrast levels were set so that the thalamus contrasted well with the caudate nucleus, lateral ventricle, third ventricle, transverse fissure, posterior limb of the internal capsule, and cerebral peduncles. Brightness and contrast levels were not changed during the segmentation of the thalamus and were identical across participants.

[1] Thalamus masks and thalamic nuclei segmentation were performed by Jon Wisco at the Massachusetts General Hospital

Table 4.1: Thalamic nuclei labels used in this thesis.

Initials	Full Name
Ant	Anterior
CM	Center Median
LD	Lateral Dorsal
LP	Lateral Posterior
LGN	Lateral Geniculate
MD	Mediodorsal
MGN	Medial Geniculate
Pu	Pulvinar
VA	Ventral Anterior
VL	Ventral Lateral
VP	Ventral Posterior

Voxels corresponding to the anterior, medial, and lateral divisions, as well as the intralaminar and reticular nuclei of the thalamus, were included in the segmentation mask. The demarcation of the borders between the reticular nucleus and the posterior limb of the internal capsule, between the ventral thalamus and zona incerta/subthalamic nucleus, and between lateral geniculate nucleus and optic radiations was viewed on all three orthogonal image planes and verified using the Nolte atlas [55].

4.5.4 Manual Nuclei Segmentation

Each hemisphere was further segmented into its nuclei on the corresponding tensor map by the neuroanatomist, following the drawings of [59]. Nuclei that could be visually identified consistently across individuals on the diffusion tensor cuboid maps were considered. The nuclei were: LGN, MGN, Pu, VA/VL/VP, CM, MD/Ant, and LD/LP. The full names for each nuclei are presented in Table 4.1. To visualize and record the position of the thalamic nuclei, an interactive labeling tool was designed by the author of this thesis to display a diffusion tensor cuboid map for each individual hemisphere overlaid on the FA images. Each voxel inside the thalamus was assigned to one of the nuclei by a trained neuroanatomist (JW). The labeled thalami were used to validate the algorithms described in this Chapter.

4.5.5 Validation

To quantify the quality of clustering results using different methods and different similarity measures, we need a measure of quality. Dice's coefficient (DC), which provides such a measure, is a similarity measure to compare two sets, A and B, and is defined as:

$$\mathrm{DC} = \frac{2 \times N(A \cap B)}{N(A) + N(B)}.$$

Dice's coefficient is used with medical images to quantify volume overlap of different segmentation results [88]. It is also possible to extend this measure to quantify volume overlaps of k-way segmentation. Using our notation, the resulting measure becomes:

$$\begin{aligned} \mathrm{VolumeOverlap} &= \frac{\sum_{i=1}^{k} 2 \times N(I_i \cap L_i)}{\sum_i N(I_i) + N(L_i)} \\ &= \frac{\sum_{i=1}^{k} N(I_i \cap L_i)}{N(L)}, \end{aligned} \tag{4.3}$$

where $L = \{L_1, L_2, \ldots, L_k\}$ are the expert labels. However, in general an unsupervised clustering algorithm, such as the k-means algorithm of [81], does not provide labels that correspond one-to-one to the manual labels. In fact, in [81] this problem is called "nuclei identification" and solved manually. The nuclei identification occurs after the clustering of the thalamus, and it refers to manually assigning anatomical labels to the clusters. During the nuclei identification, merging of clusters is allowed, which results through assigning the same anatomical label to multiple clusters.

In more general terms, the nuclei identification problem is similar to the correspondence problem of computer vision, which can be formulated as the following: Given two sets of indexes $I = \{I_1, I_2, \ldots, I_k\}$, containing clustering results and $L = \{L_1, L_2, \ldots, L_m\}$ containing expert labels, the correspondence mapping is a set of pairs (i, l), where $i \in I$ and $l \in L$. For each pair of assignments, there is an affinity $A(i, j)$ which quantifies the cost of making this assignment. The number of clusters k is not required to be the same as m, and in the case $k > m$, the solution to the nuclei identification problem involves merging clustering results as in [81]. For our specific problem, there are exactly k pairs in the

correspondence mapping, since each cluster can only be assigned to a unique anatomical label.

To solve this problem automatically, we need to identify a cost measure for each possible pair of assignments. Since we are quantifying the segmentation quality using the volume overlap measure as defined in Equation 4.3, it is easy to show that the corresponding cost function is $N(I_i \cap L_l)$. Therefore, to obtain the maximum volume overlap, each clustering result needs to be assigned to the anatomical label. Such an assignment results in the maximum number of overlapping voxels. We note that as the number of clusters k increases, the volume overlap measure calculated this way is guaranteed to increase. When k is set to be equal to the number of voxels, the solution to the problem of nuclei identification becomes equivalent to manual segmentation. We will discuss this point further in the conclusion of this chapter, as the previous work uses higher values of k without noting this subtlety.

4.6 Experiments

4.6.1 Comparison of Heuristic Solutions to Normalized Cuts

We applied three spectral clustering algorithms to in-vivo DTI data from ten subjects. The three algorithms used were

1. the spectral splitting as described in [67],
2. spectral split and merge as described in Section 4.3,
3. spectral split and merge followed by object swaps.

f_{ang} was used for tensor dissimilarity and the same edge-weight matrix was used for all three algorithms. To demonstrate the qualitative differences between the three algorithms, a single slice from an exemplary subject that was segmented by the three algorithms is shown in Figure 4-4 along with the resulting normalized cut value for the respective segmentations. As the normalized cut value decreases, the boundaries between the segments follow more closely the valleys of low edge-weights as shown in the corresponding graph.

Normalized cut values along with volume overlaps of the resulting segmentations with expert labeled data is presented in Figure 4-5 for all ten subjects. As in Figure 4-4, the lowest normalized cut value is achieved with spectral split and merge, followed by object swaps, and the highest values were obtained through the spectral splitting as described in [67]. Furthermore, the lower normalized cut values translated into higher volume overlaps with the expert labels.

4.6.2 Comparison of Tensor Similarity Measures

We applied the spectral split and merge algorithm, followed by object swaps, to in-vivo DTI data from ten subjects using the three tensor similarity measures of Section 4.4.1. We also repeated the same experiment with the k-means algorithm as described in Section 4.2.1 to use the k-means algorithm as a benchmark. Volume overlaps with expert labels are presented for these experiments in Figure 4-6 for $k = 7$ and in Figure 4-7 for $k = 12$. In both figures the results are presented separately for the left and right hemispheres. The

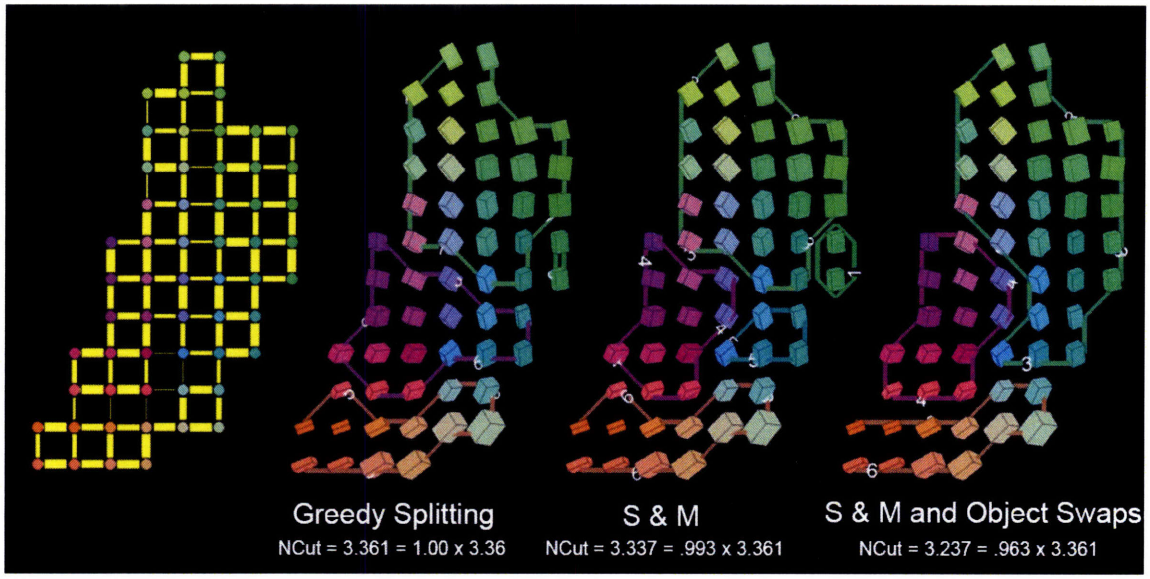

Figure 4-4: A single slice from an exemplary subject segmented by three spectral algorithms is shown. f_{ang} was used for tensor dissimilarity and the same edge-weight matrix was used for all three algorithms. Even though only a single slice is shown, the segmentation was done in 3D for the whole thalamus. The left-most figure shows the corresponding graph, with each vertex of the graph (voxel) represented by a colored circle with the edge-weights reflected by the thickness of the edge connecting the vertices. Only the edges for the face-neighbors are shown for clarity. The resulting normalized cut values are as indicated under each segmented slice.

resulting error bars are similar for both hemispheres. The spectral algorithms result in higher volume overlaps as opposed to the k-means algorithm. As for the tensor similarity measures, f_{ang} results in the highest volume overlaps, whereas the f_{KL} results in the lowest. Volume overlaps are higher with $k = 12$ as opposed to the results with $k = 7$.

4.6.3 Qualitative Results from Spectral Clustering on DTI Data

For all thalamic data sets, spectral reordering of the edge-weight matrix revealed a significant clustering structure (Figure 4-8D). The graph created by the algorithm provides necessary information for the clustering. Furthermore, the spectral ordering followed by the modified spectral clustering algorithm was able to identify the clusters. These clusters are presented in Figure 4-9 (Right) for one of the individuals along with the expert labels (Left). Automatically segmented clusters are similar in both hemispheres, which were pro-

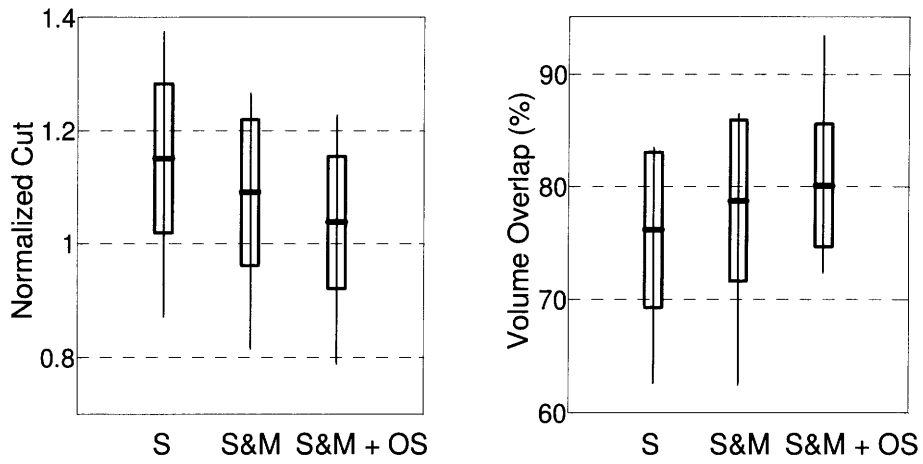

Figure 4-5: Normalized cut values along with volume overlaps with expert labels from three spectral clustering methods are presented for 10 subjects. S stands for spectral splitting, S&M stands for split and merge and S&M+OS stands for split and merge followed by object swaps. The error bars show the mean value of the respective quantity as well as the range within one standard deviation calculated from the experimental results. The thin lines indicate the range of values obtained in all experiments.

cessed independently, and they match well with the expert segmentation. The clustering was achieved by the split and merge algorithm, followed by object swaps, and the resulting hierarchy tree is presented in Figure 4-10.

4.6.4 Effect of k

To investigate the effect of k on the segmentation quality, DTI data from ten subjects were segmented with varying number of clusters $k \in \{5, 10, \ldots, 50\}$. The volume overlap measure increased monotonically with increasing k as shown in Figure 4-11. There are around 1000 voxels in each thalamus, therefore, the volume overlap would have reached to 100% as k reached the number of voxels.

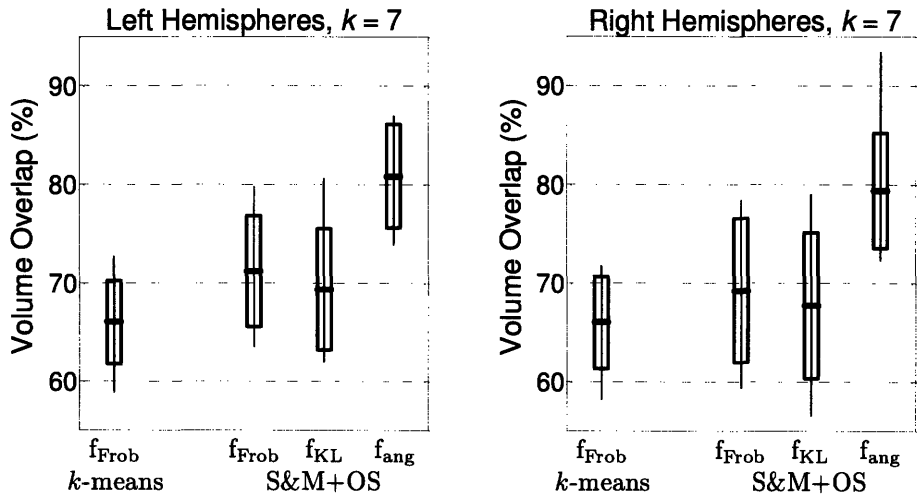

Figure 4-6: Volume overlaps between the expert labeled data and clustering results with different tensor similarity measures for the left and right hemispheres of 10 subjects. The number of clusters were set to 7 in all experiments. k-means and the spectral split and merge, followed by object swaps, were used for clustering.

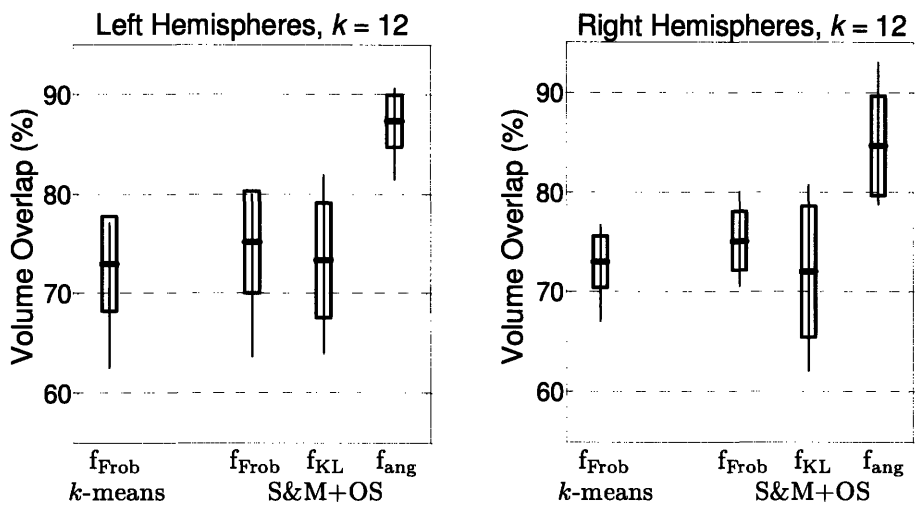

Figure 4-7: Volume overlaps between the expert labeled data and clustering results with different tensor similarity measures for the left and right hemispheres of 10 subjects. The number of clusters were set to 12 in all experiments. k-means and the spectral split and merge, followed by object swaps, were used for clustering.

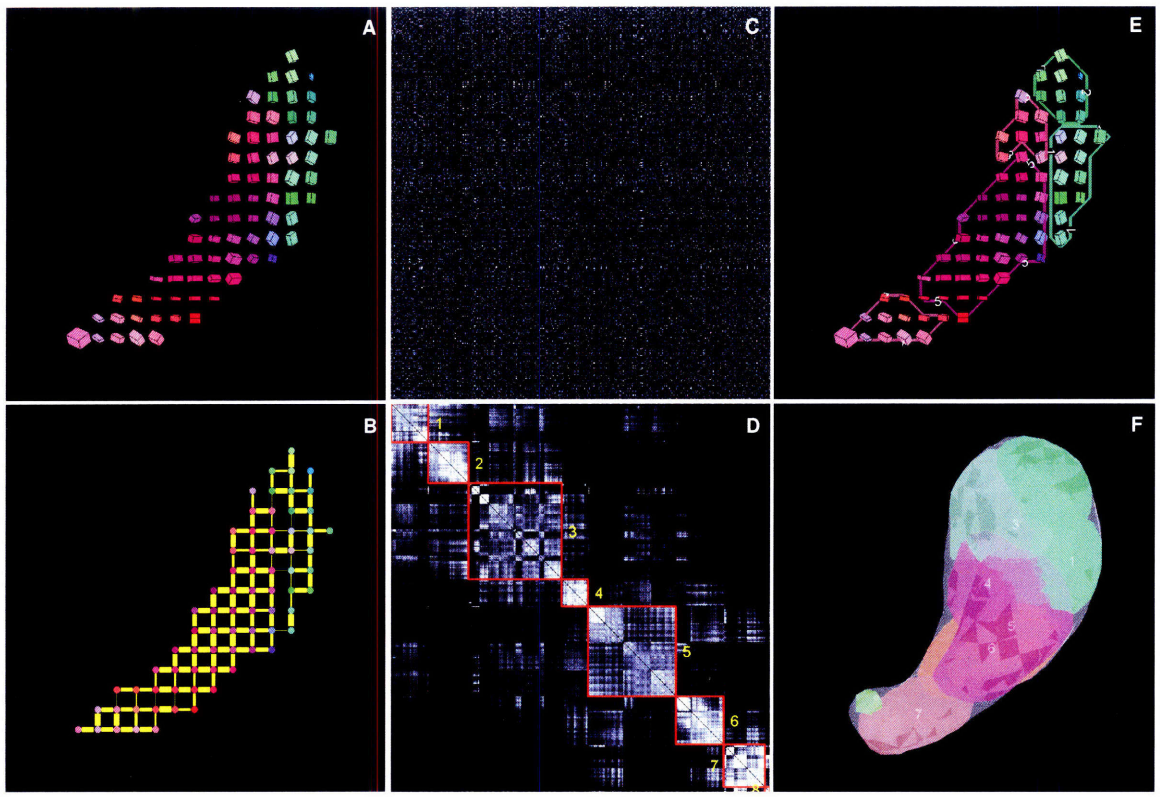

Figure 4-8: A schematic outline of spectral segmentation algorithm. (A) A single slice tensor data and (B) the corresponding graph with each vertex of the graph (voxel) represented by a colored circle with the edge-weights reflected by the thickness of the edge connecting the vertices. Only the edges for the face-neighbors are shown for clarity. (C) Edge-weight matrix $\mathbf{W}$ with initial random ordering did not show clustering structure, but (D) the spectral ordering presents clustering structure. The resulting clusters are marked by the red contours. (E) The resulting segmentation is presented for one slice, (F) as well as the for the whole thalamus in 3D.

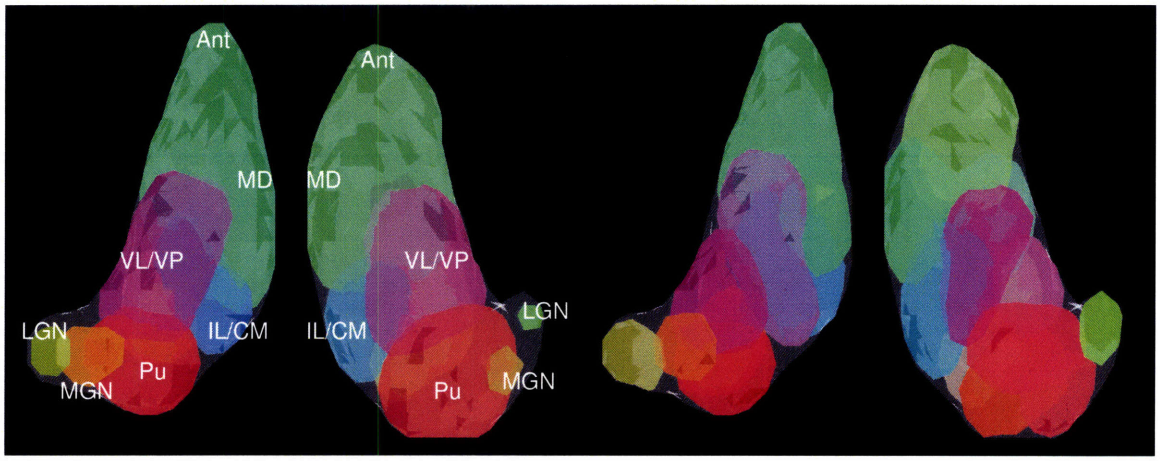

Figure 4-9: Left: 3D rendering of expert segmentation of both hemispheres from one subject. Each cluster is colored according to the mean diffusion direction of the cluster. Right: The same subject segmented by the modified spectral clustering with $k = 12$.

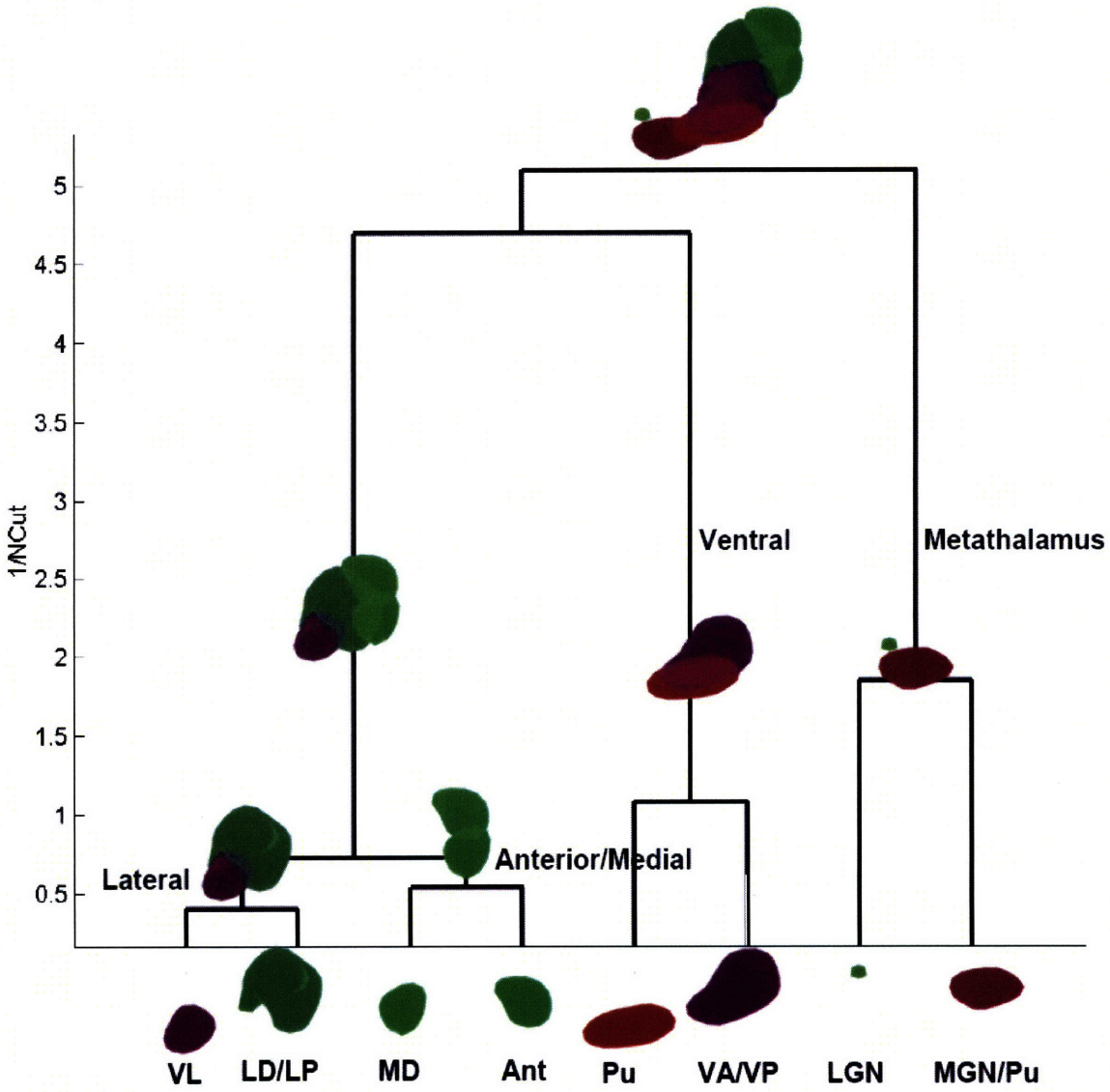

Figure 4-10: An example hierarchy tree obtained through spectral clustering.

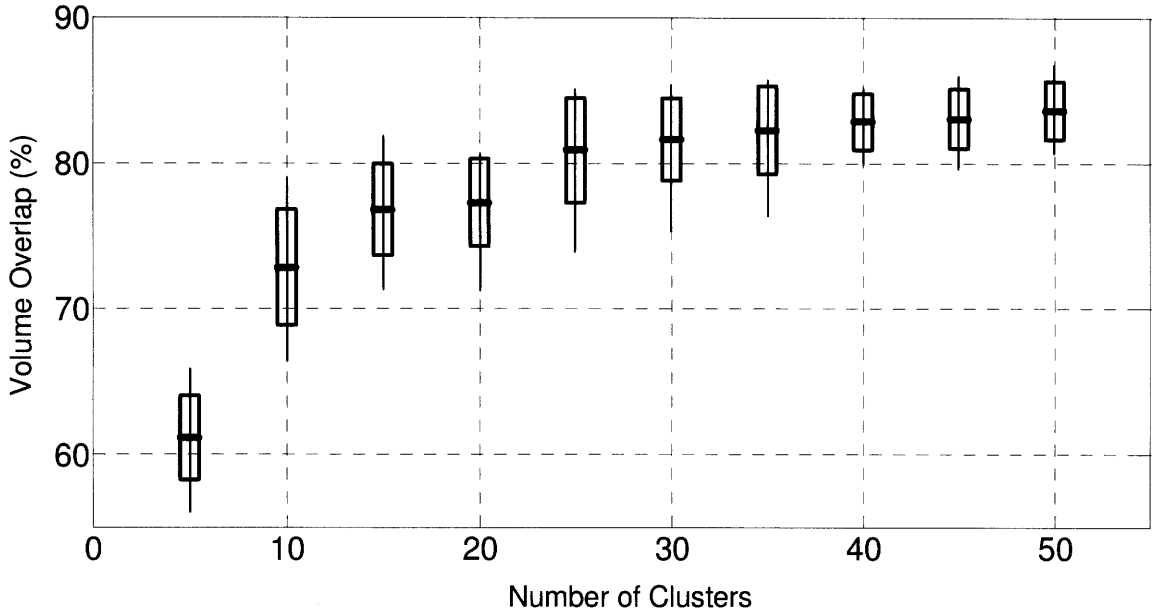

Figure 4-11: Volume overlaps between the expert labeled data and k-means clustering results with differing k. As k becomes closer to the number of voxels present in the thalamus, the overlap measure will asymptotically approach to 100%.

4.7 Summary of the Chapter

In this section, we investigated several clustering algorithms with different tensor similarity metrics for segmentation of thalamic nuclei. All of the algorithms investigated were able to produce plausible segmentations for the thalami. However, the spectral split and merge clustering, followed by the object swaps, produced both qualitatively and quantitatively better results when compared to the other algorithms. Furthermore, all of the spectral methods investigated in this chapter worked better than the previously proposed k-means algorithm.

Another variable for varying performance is the choice of tensor similarity metric used by the clustering algorithms. The full-tensor-based metrics such as the Frobenius norm and KL-divergence failed to match the accuracy of segmentations when compared to our metric that depends only on the principal diffusion direction. We believe the reason for that is due to the cellular properties of the thalamus. The thalamic nuclei consist mainly of gray matter structure with little mylenation and directional coherence due to the major white matter tracts connecting to the cortex, which in return results in nearly isotropic voxels. The full-tensor-based metrics quantify mostly the similarity due to the largely isotropic tensors and to a smaller degree the directional information. However, the principal diffusion direction disregards the isotropic part of the tensors and only quantify the orientation information. This could result in a noisier metric; however, our results indicate that the advantage of disregarding non-informative isotropic parts of the tensors outweigh the effects of noise.

A promising result is the negative correlation between the normalized cut value and the volume overlap measure observed in all the experiments. This result indicates that the normalized cut criteria is a good metric for segmentation of diffusion tensor data.

One major problem of segmentation through clustering individual thalami is the lack of correspondence between clusters obtained in different subjects. This results in an inability of assigning anatomical labels automatically. This problem has been addressed previously through expert identification of the resulting clusters in each subject. In this chapter, we solve this problem using an automatic assignment process that maximizes the volume overlap measure between the clustering results and the expert labels. We note that this is not

a suitable way to assign anatomical labels for unlabeled thalami; however, it was suitable for our objective of quantifying the volume overlaps. An interesting result of this kind of anatomical labeling is that as the number of clusters k increases, the volume overlap measure is guaranteed to increase. And we confirmed this in our experimental setting in Figure 4-11. Both in [81] and [33] k was set at 14. At that level of k with expert identification of anatomical labels, it becomes possible to obtain high volume overlaps between subjects; however, this kind of user interaction makes the resulting segmentations less automatic.

Chapter 5

Consistent Gray Matter Segmentation

In Chapter 4, we investigated several unsupervised clustering algorithms for segmenting thalamic nuclei from diffusion tensor images. These algorithms produce a plausible segmentation on individual subjects; however, they do not address the problem of consistently identifying the same functional areas in a population. The lack of correspondence between the segmented nuclei makes it more difficult to use the results from the unsupervised segmentation tools for morphometry. In this Chapter we present a novel segmentation algorithm to automatically segment the gray matter nuclei while ensuring consistency between subjects in a population. This new algorithm, referred to as Consistency Clustering, finds correspondence between the nuclei during the segmentation since the segmentation is achieved through a single model for the whole population. This model is an output of the algorithm and it serves as a probabilistic atlas similar to the brain atlases experts use to identify thalamic nuclei. Furthermore, the Consistency Clustering can utilize both previously labeled and unlabeled data, therefore the algorithm can be used both as an unsupervised or as a semi-supervised method for creating probabilistic brain atlases.

5.1 Introduction

Diffusion tensor imaging (DTI) is a relatively new imaging modality that measures free water diffusion, i.e. Brownian motion, of the endogenous water in tissue [6]. In human brain tissue, the water diffusion is not the same in all directions, since it is obstructed by structural elements such as cell membranes or myelin [6]. When this obstruction constrains the water diffusion in a coherent direction, such as within the cerebral white matter, the resulting water diffusion tensor becomes anisotropic, containing information about the directionality of the white matter connectivity. Thus, quantification of water diffusion in tissue through DTI provides a unique way to analyze white matter organization of the brain.

Unlike white matter, the tissue in gray matter is less organized in orientation. The lack of coherent orientation limits the use of DTI for gray matter analysis in some areas, such as the cerebral cortex. However, there are certain gray matter structures that exhibit coherence in diffusion direction due to the presence of coherent white matter near these structures, such as the thalamus. The thalamus acts as the central relay station of the brain with nearly all of the sensory tracts project to the cortex passing through the thalamus. Since functionally related pathways target the same region of cortex once they leave the thalamus, they result in organization of diffusivity within the thalamus. This organized diffusion can be measured in DTI, and it has been proposed that the thalamic nuclei can be distinguished by their characteristic diffusion orientation [81].

Precise identification of the thalamic nuclei is essential in a clinical setting, since many motor-control disorders are surgically corrected by applying chronic electrical stimulation to the appropriate functional area of the thalamus. Currently, these regions are detected qualitatively before a surgical operation using generic atlases along with structural MRI [26], even though the structural MRI does not provide adequate contrast to identify the distinct nuclei. Changes have also been reported in the thalamic nuclei during the progression of a large number of diseases, including schizophrenia [63] and Parkinson's disease [24]

Using the knowledge that thalamic nuclei can be resolved through DTI, several segmentation algorithms have been proposed to segment the thalamic nuclei. The earliest segmentation method, which depends on DTI data only from within the thalamus, uses

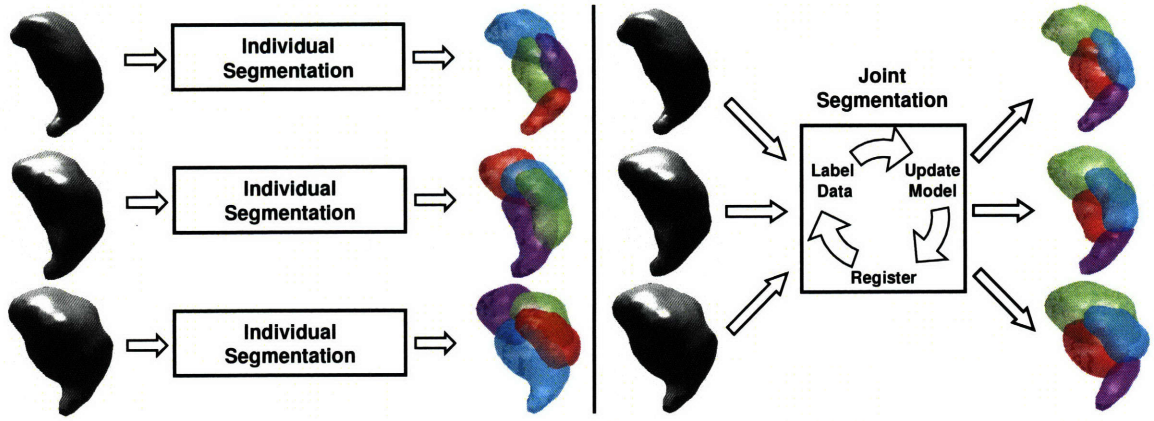

Figure 5-1: Schematic description of previous thalamus segmentation algorithms [81, 86, 32, 19] (left) as opposed to the Consistency Clustering (right).

the k-means clustering algorithm [81]. Other clustering methods have been proposed that use spectral clustering [86], level-sets [32] and the mean-shift algorithm [19]. These other methods avoid some of the weaknesses of k-means, which includes a bias toward ellipsoidal clusters and sensitivity to initialization. Even though each of these clustering algorithms produce plausible segmentations for any given subject in a population, they do not find a correspondence between the segments acquired from different subjects.

In this Chapter, we present a new approach to the segmentation of thalamic nuclei. Unlike the previous methods, this new algorithm, referred to as Consistency Clustering (CC), is designed to segment multiple subjects simultaneously and find a correspondence between the segmentation results (Figure 5-1). The CC achieves these goals by learning a thalamic model of the population under investigation, which serves as a probabilistic atlas of the thalamic nuclei. This model involves a spatial component as well as a directional component for each nuclei. The CC also performs a non-linear (poly-rigid) registration to account for inter-subject variability. Since the segmentation of each individual subject is done according to a common model, the consistency of segmentations between subjects is ensured. We formulate the joint segmentation problem as a maximum likelihood problem which the Consistency Clustering solves using a generalized Expectation Maximization (gEM) framework. This joint segmentation approach results in a segmentation for each subject and determines a correspondence between subjects. Also, the thalamic model,

which is learned from the population, serves as a compact representation of the population under investigation, which finds applications in quantitative morphometry.

In the following Sections, we first provide background on the expectation maximization (EM) algorithm, and then describe the theory behind the method which uses EM for optimization. Finally, we present results from several experiments that demonstrate the feasibility of the proposed method with DTI data from 10 healthy participants.

5.2 Background on Expectation Maximization

Expectation maximization (EM) is a numerical algorithm for optimization of parametric functions. In this section we will provide a brief background on the EM algorithm and its application to clustering through Gaussian and von-Mises Fisher mixture models.

5.2.1 An Iterative Lower Bound Optimization

The intuition behind EM is a simple one: Given a function $J(\Theta)$, whose value we wish to maximize, and an initial estimate of the parameter set $\Theta^{(0)}$, the EM algorithm iteratively updates the parameter set $\Theta^{(n)}$ until $J(\Theta)$ can no longer be increased. EM achieves this task by identifying a lower bound $b_n(\Theta)$, at every iteration (n), which satisfies the following two conditions:

- b_n is a lower bound for J; i.e., $b_n(\Theta) \leq J(\Theta)$ for all Θ,
- b_n touches J at the $(n)^{th}$ estimate of the parameter set; i.e., $b_n(\Theta^{(n)}) = J(\Theta^{(n)})$.

An illustration of these conditions is provided in Figure 5-2. Given these two conditions, it is clear that the $(n+1)^{st}$ parameter set, which is chosen to maximize the lower bound $b_n(\Theta)$, also results in a higher or equal value of the optimization function; i.e. $J(\Theta^{(n+1)}) \geq J(\Theta^{(n)})$. Through this observation, the EM algorithm reduces the problem of maximizing $J(\Theta)$ to identifying a lower bound function that is easier to maximize. Identification of such a lower bound is not a trivial problem, but good choices for such a lower bound exist for some important family of functions, such as the parametric density estimators. We refer the reader to [51] for formal proofs of these statements and for a large set of problems with known solutions.

5.2.2 Maximum Likelihood Density Estimation

We define a maximum likelihood density estimation problem as following: Given a set of N points in D dimensions, $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \,|\mathbf{x}_i \in \mathbb{R}^D\}$, and a family $\mathscr{G}$ of density probability density functions defined on $\mathbb{R}^D$, calculate the probability density function

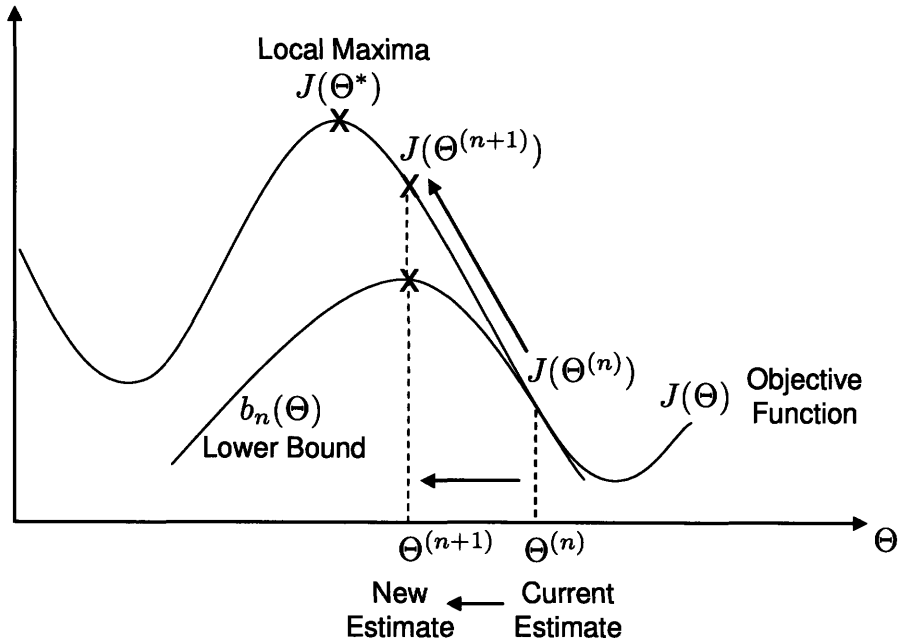

Figure 5-2: Illustration of an EM iteration. The algorithm identifies a lower bound $b_n(\Theta)$ for the $(n)^{th}$ estimate of the parameter set $\Theta^{(n)}$, which touches the objective function $J(\Theta)$ at $\Theta^{(n)}$. Calculating the maximum of the lower bound in some cases is easier then identifying the maximum of the objective function. Iteration (n) is completed by obtaining the parameter set that maximizes the lower bound. The algorithm is guaranteed to find the local maxima of $J(\Theta)$ through these iterations.

$p(\mathbf{x};\Theta) \in \mathscr{G}$ that maximizes the likelihood of the observations in X. Since the density functions in this family are parametrized with Θ, this problem is equivalent with finding Θ^* that maximizes the likelihood of observations in X. For an arbitrary density function family, this problem is formalized as:

$$\begin{aligned}\Theta^* &= \arg\max_{\Theta} L(\Theta|X), \\ &= \arg\max_{\Theta} \prod_{i=1}^{N} p(\mathbf{x}_i;\Theta),\end{aligned}$$

where the samples $\mathbf{x}_i$ are assumed to be independent. It is common practice to use a monotonic function, such as the logarithm function, to linearize a problem that involves

products. In that case, our density estimation problem becomes:

$$\begin{aligned}
\Theta^* &= \arg\max_{\Theta} \log L(\Theta|X), \\
&= \arg\max_{\Theta} \mathscr{L}(\Theta|X), \\
&= \arg\max_{\Theta} \sum_{i=1}^{N} \log p(\mathbf{x}_i; \Theta).
\end{aligned}$$

Here $\mathscr{L}(\Theta)$ denotes the log-likelihood and it is analogous to the optimization function $J(\Theta)$ of Section 5.2.1.

Class Labels

We now introduce class labels $c \in \{1, 2, \ldots, C\}$, which serves as the mixture weights in our problem:

$$\begin{aligned}
\mathscr{L}(\Theta) &= \sum_{i=1}^{N} \log \sum_{c=1}^{C} p(\mathbf{x}_i, c; \Theta), \qquad (5.1) \\
&= \sum_{i=1}^{N} \log \sum_{c=1}^{C} p(c; \Theta) p(\mathbf{x}_i | c; \Theta), \\
&= \sum_{i=1}^{N} \log \sum_{c=1}^{C} \pi_c p(\mathbf{x}_i | c; \Theta),
\end{aligned}$$

where $\pi_c = p(c; \Theta)$ is the mixing weight for mixture (class) c. To ensure $p(c; \Theta)$ is a valid probability distribution, the mixture weights need to be non-negative and sum up to one; i.e.,

$$\begin{aligned}
\pi_c &\geq 0, \\
\sum_c \pi_c &= 1.
\end{aligned}$$

Introduction of the class labels enables finding a lower bound $b_n(\Theta)$. Class labels are usually called "hidden variables" since they are not observed in the random experiment, but they appear in the optimization function as variables.

Jensen's Inequality

We use Jensen's inequality to construct a lower bound for the log-likelihood in our problem. Jensen's inequality provides a lower bound for any real concave function φ,

$$\varphi\left(\frac{\sum_i \alpha_i y_i}{\sum_i \alpha_i}\right) \geq \frac{\sum_i \alpha_i \varphi(y_i)}{\sum_i \alpha_i}, \tag{5.2}$$

where $\{y_i\}$ denote samples in the domain of φ and $\{\alpha_i\}$ denote non-negative weights for the corresponding samples. Adding an additional constraint on the weights, $\sum_i \alpha_i = 1$ and choosing $\log$ as the concave function, Jensen's inequality becomes:

$$\log \sum_i \alpha_i y_i \geq \sum_i \alpha_i \log(y_i). \tag{5.3}$$

Lower Bound

To use Jensen's inequality to construct a lower bound in our problem, we revisit the log-likelihood from Equation 5.1, and introduce a generic probability distribution $q(c)$ defined over the class labels:

$$\begin{aligned} \mathscr{L}(\Theta) &= \sum_{i=1}^{N} \log \sum_{c=1}^{C} p(\mathbf{x}_i, c; \Theta), \\ &= \sum_{i=1}^{N} \log \sum_{c=1}^{C} p(c; \Theta) p(\mathbf{x}_i | c; \Theta) \frac{q(c)}{q(c)}, \end{aligned} \tag{5.4}$$

where $\sum_c q(c) = 1$. Then, using the Jensen's inequality from Equation 5.3, we obtain a lower bound for the log-likelihood:

$$\mathscr{L}(\Theta) = \sum_{i=1}^{N} \log \sum_{c=1}^{C} p(\mathbf{x}_i, c; \Theta) \frac{q(c)}{q(c)} \geq \sum_{i=1}^{N} \sum_{c=1}^{C} q(c) \log\left(\frac{p(\mathbf{x}_i, c; \Theta)}{q(c)}\right), \tag{5.5}$$

where we substituted $q(c)$ for α_i and $(p(\mathbf{x}_i, c; \Theta)/q(c))$ for y_i in Equation 5.3. Further manipulation of this lower bound results in:

$$\begin{aligned}
\sum_{i=1}^{N}\sum_{c=1}^{C} q(c)\log\left(\frac{p(\mathbf{x}_i, c; \Theta)}{q(c)}\right) &= \sum_{i=1}^{N}\sum_{c=1}^{C} q(c)\log p(\mathbf{x}_i, c; \Theta) - \sum_{i=1}^{N}\sum_{c=1}^{C} q(c)\log q(c), \\
&= \sum_{i=1}^{N}\sum_{c=1}^{C} q(c)\log p(\mathbf{x}_i, c; \Theta) - \text{Constant}, \\
&= \sum_{i=1}^{N} E_q[\log p(\mathbf{x}_i, c; \Theta)] - \text{Constant},
\end{aligned}$$

where $E_q[.]$ denotes an expectation with respect to q. The appearance of the expectation in the lower bound is the reason for the name of expectation maximization algorithm.

In Equation 5.5, any choice for function $q(c)$, from the set of valid distributions over c, is guaranteed to be a lower bound for the log-likelihood; however, we also need to ensure that the resulting lower bound touches the log-likelihood at $\Theta^{(n)}$. This second condition is met for $q(c) = p(c|\mathbf{x}_i; \Theta^{(n)})$, which is easy to see by plugging in $p(c|\mathbf{x}_i; \Theta^{(n)})$ for $q(c)$ and setting $\Theta = \Theta^{(n)}$ in the lower bound from Equation 5.5:

$$\begin{aligned}
\sum_{i=1}^{N}\sum_{c=1}^{C} q(c)\log\left(\frac{p(\mathbf{x}_i, c; \Theta)}{q(c)}\right) &= \sum_{i=1}^{N}\sum_{c=1}^{C} p(c|\mathbf{x}_i; \Theta^{(n)})\log\left(\frac{p(\mathbf{x}_i, c; \Theta^{(n)})}{p(c|\mathbf{x}_i; \Theta^{(n)})}\right), \\
&= \sum_{i=1}^{N}\sum_{c=1}^{C} p(c|\mathbf{x}_i; \Theta^{(n)})\log\left(\frac{p(\mathbf{x}_i, c; \Theta^{(n)})p(\mathbf{x}_i; \Theta^{(n)})}{p(c|\mathbf{x}_i; \Theta^{(n)})p(\mathbf{x}_i; \Theta^{(n)})}\right), \\
&= \sum_{i=1}^{N}\sum_{c=1}^{C} p(c|\mathbf{x}_i; \Theta^{(n)})\log\left(\frac{p(\mathbf{x}_i, c; \Theta^{(n)})p(\mathbf{x}_i; \Theta^{(n)})}{p(\mathbf{x}_i, c; \Theta^{(n)})}\right), \\
&= \sum_{i=1}^{N}\sum_{c=1}^{C} p(c|\mathbf{x}_i; \Theta^{(n)})\log\left(p(\mathbf{x}_i; \Theta^{(n)})\right), \\
&= \sum_{i=1}^{N}\log\left(p(\mathbf{x}_i; \Theta^{(n)})\right)\sum_{c=1}^{C} p(c|\mathbf{x}_i; \Theta^{(n)}), \\
&= \sum_{i=1}^{N}\log\left(p(\mathbf{x}_i; \Theta^{(n)})\right), \\
&= \mathscr{L}(\Theta^{(n)}),
\end{aligned}$$

where we used the fact that $\sum_{c=1}^{C} p(c|\mathbf{x}_i; \Theta^{(n)}) = 1$. The resulting lower bound $b_n(\Theta)$ is

written explicitly as:

$$b_n(\Theta) = \sum_{i=1}^{N}\sum_{c=1}^{C} p(c|\mathbf{x}_i;\Theta^{(n)}) \log p(\mathbf{x}_i, c;\Theta) - \sum_{i=1}^{N}\sum_{c=1}^{C} p(c|\mathbf{x}_i;\Theta^{(n)}) \log p(c|\mathbf{x}_i;\Theta^{(n)}),$$

where the second summation does not depend on Θ and therefore maximizing $b_n(\Theta)$ is equivalent to maximizing $\widetilde{b}_n(\Theta)$:

$$\widetilde{b}_n(\Theta) = \sum_{i=1}^{N}\sum_{c=1}^{C} p(c|\mathbf{x}_i;\Theta^{(n)}) \log p(\mathbf{x}_i, c;\Theta). \tag{5.6}$$

The EM algorithm maximizes this lower bound iteratively, with two stages per iteration. In the first stage, called the expectation step or "E-Step", the algorithm constructs the lower bound $\widetilde{b}_n(\Theta)$ using the current estimate of the parameter set, $\Theta^{(n)}$. In the second stage, called the maximization step or "M-Step", the algorithm maximizes the bound $\widetilde{b}_n(\Theta)$ by calculating the parameter set $\Theta^{(n+1)}$ that maximizes the lower bound (See Figure 5-2).

5.2.3 EM for Estimating a Gaussian Mixture Model

We now derive the specific update equations for estimating a Gaussian mixture model using the expectation maximization algorithm. In Section 5.2.2, we define the density estimation problem as following: Given a set of N points in D dimensions, $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \,|\mathbf{x}_i \in \mathbb{R}^D\}$, and a family $\mathscr{G}$ of density probability density functions defined on $\mathbb{R}^D$, calculate the probability density function $g(\mathbf{x}) \in \mathscr{G}$ that maximizes the likelihood of the observations in X. For a Gaussian mixture model, the density functions are conditioned to be mixtures of Gaussians, i.e.,

$$g(\mathbf{x}) = \sum_{c=1}^{C} \pi_c f(\mathbf{x}|c;\Theta)$$

where π_c is the mixing weight (probability) for mixture c. $f(\mathbf{x}|c;\Theta)$ is a Gaussian, i.e.

$$f(\mathbf{x}|c;\Theta) = N(\mathbf{x};\mu_c,\mathbf{\Sigma}_c) = \frac{1}{(2\pi)^{3/2}|\mathbf{\Sigma}_c|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_c)^T\mathbf{\Sigma}_c^{-1}(\mathbf{x}-\mu_c)\right),$$

where μ_c is the mean vector and $\mathbf{\Sigma}_c$ is the covariance matrix for mixture c. As before, the mixture weights need to sum up to one, $\sum_c \pi_c = 1$ and be non-negative, $\pi_c \geq 0$. Given this model, the log-likelihood of X becomes:

$$\begin{aligned}
\mathscr{L}(\Theta) &= \log L(\Theta), \\
&= \log \prod_{i=1}^{N} \sum_{c=1}^{C} \pi_c N(\mathbf{x}_i; \mu_c, \mathbf{\Sigma}_c), \\
&= \sum_{i=1}^{N} \log \sum_{c=1}^{C} \pi_c N(\mathbf{x}_i; \mu_c, \mathbf{\Sigma}_c).
\end{aligned}$$

Assuming we start with an initial estimate of the parameters $\Theta^{(0)} = \{\Theta_1^{(0)}, \Theta_2^{(0)}, \ldots, \Theta_C^{(0)}\}$, where $\Theta_c^{(0)} = \{\pi_c^{(0)}, \mu_c^{(0)}, \mathbf{\Sigma}_c^{(0)}\}$, the expectation maximization algorithm calculates a better estimate for the parameter set first by constructing a lower bound $b_n(\Theta)$ in the E-Step and then maximizing this bound with respect to the parameters in Θ in the M-Step. We now derive the specific update equations for the Gaussian mixture model that is used to iteratively estimate Θ^*.

E-Step

In the E-Step the algorithm constructs the lower bound $b_n(\Theta)$ by taking the expectation of the log-likelihood with respect to the current distribution of the class labels. From Equation 5.6, the lower bound (excluding the constant term) is:

$$\begin{aligned}
\widetilde{b}_n(\Theta) &= \sum_{i=1}^{N} \sum_{c=1}^{C} p(c|\mathbf{x}_i; \Theta^{(n)}) \log p(\mathbf{x}_i, c; \Theta), \\
&= \sum_{i=1}^{N} \sum_{c=1}^{C} p(c|\mathbf{x}_i; \Theta^{(n)}) \log \pi_c N(\mathbf{x}_i; \mu_c, \mathbf{\Sigma}_c), \qquad (5.7)
\end{aligned}$$

where we plugged in the specific equations for Gaussian mixture model in the second step. To construct this bound, we need to calculate $p(c|\mathbf{x}_i; \Theta^{(n)})$, which are called "membership

probabilities", for each of the samples in X:

$$\begin{aligned} p(c|\mathbf{x}_i;\Theta^{(n)}) &= \frac{p(\mathbf{x}_i|c;\Theta^{(n)})p(c;\Theta^{(n)})}{p(\mathbf{x}_i;\Theta^{(n)})}, \\ &= \frac{\pi_c^{(n)} N(\mathbf{x}_i;\mu_c^{(n)},\mathbf{\Sigma}_c^{(n)})}{\sum_{c=1}^{C}\pi_c^{(n)} N(\mathbf{x}_i;\mu_c^{(n)},\mathbf{\Sigma}_c^{(n)})}, \\ &\triangleq p_{ci}^{(n)}, \end{aligned} \tag{5.8}$$

where $p_{ci}^{(n)}$ denote the membership probabilities.

M-Step

In the M-Step, the algorithm updates the parameter set Θ to maximize the lower bound $b_n(\Theta)$. Update equations for each of the parameters are derived using Lagrange multipliers for the corresponding constraints (if any) and setting the derivative of the lower bound to zero. We first derive the update equation for π_c with the constraint $\sum_c \pi_c = 1$:

$$\begin{aligned} H_\pi(\Theta) &= \sum_{i=1}^{N}\sum_{c=1}^{C} p_{ci}^{(n)} \log \pi_c N(\mathbf{x}_i;\mu_c,\mathbf{\Sigma}_c) - \lambda\left(\sum_{c=1}^{C}\pi_c - 1\right), \\ &= \sum_{i=1}^{N}\sum_{c=1}^{C} p_{ci}^{(n)} \log \pi_c - \lambda\left(\sum_{c=1}^{C}\pi_c - 1\right) - \sum_{i=1}^{N}\sum_{c=1}^{C} p_{ci}^{(n)} \log N(\mathbf{x}_i;\mu_c,\mathbf{\Sigma}_c). \end{aligned}$$

Taking the derivative of $H_\pi(\Theta)$ with respect to a specific mixture weight, π_j for $c = j$ results in:

$$\frac{\partial H_\pi(\Theta)}{\partial \pi_j} = \sum_{i=1}^{N} \frac{p_{ji}^{(n)}}{\pi_j} - \lambda. \tag{5.9}$$

Rearranging to solve for π_j gives us:

$$\pi_j = \frac{1}{\lambda}\sum_{i=1}^{N} p_{ji}^{(n)}.$$

To solve for λ, we sum each side over j:

$$\sum_{j=1}^{C} \pi_j = \frac{1}{\lambda} \sum_{i=1}^{N} \sum_{j=1}^{C} p_{ji}^{(n)}.$$

Since $\sum_{j=1}^{C} \pi_j = 1$, λ equals:

$$\begin{aligned} \lambda &= \sum_{i=1}^{N} \sum_{c=1}^{C} p_{ci}^{(n)}, \\ &= \sum_{i=1}^{N} 1, \\ &= N, \end{aligned}$$

where we used the equality $\sum_{c=1}^{C} p_{ci}^{(n)} = 1$. Finally, the resulting update equation for $\pi_c^{(n+1)}$ is:

$$\pi_c^{(n+1)} = \frac{1}{N} \sum_{i=1}^{N} p_{ci}^{(n)}.$$

There are no constraints on the mean vectors μ_c; therefore we do not need a Lagrange multiplier for the corresponding update equation:

$$\begin{aligned} H_\mu(\Theta) &= \sum_{i=1}^{N} \sum_{c=1}^{C} p_{ci}^{(n)} \log \pi_c N(\mathbf{x}_i; \mu_c, \mathbf{\Sigma}_c), \\ &= \sum_{i=1}^{N} \sum_{c=1}^{C} p_{ci}^{(n)} \log \pi_c + \sum_{i=1}^{N} \sum_{c=1}^{C} p_{ci}^{(n)} \log N(\mathbf{x}_i; \mu_c, \mathbf{\Sigma}_c), \\ &= \sum_{i=1}^{N} \sum_{c=1}^{C} p_{ci}^{(n)} \log \pi_c - \sum_{i=1}^{N} \sum_{c=1}^{C} \frac{1}{2} (\mathbf{x}_i - \mu_c)^T \mathbf{\Sigma}_c^{-1} (\mathbf{x}_i - \mu_c) \\ &\quad + \sum_{i=1}^{N} \sum_{c=1}^{C} \pi_c \log((2\pi)^{3/2} |\mathbf{\Sigma}_c|^{1/2}). \end{aligned}$$

Taking the derivative of $H_\mu(\Theta)$ with respect to a specific mean vector, μ_j for $c = j$ results in:

$$\frac{\partial H_\mu(\Theta)}{\partial \mu_j} = \sum_{i=1}^{N} p_{ji}^{(n)} \mathbf{\Sigma}_j^{-1}(\mathbf{x}_i - \mu_j).$$

We now set the derivative to zero:

$$\mathbf{\Sigma}_j^{-1} \sum_{i=1}^{N} p_{ji}^{(n)} \mathbf{x}_i - \mathbf{\Sigma}_j^{-1} \mu_j \sum_{i=1}^{N} p_{ji}^{(n)} = 0.$$

Rearranging to solve for μ_j results in the following update equation:

$$\mu_c^{(n+1)} = \frac{\sum_{i=1}^{N} p_{ci}^{(n)} \mathbf{x}_i}{\sum_{i=1}^{N} p_{ci}^{(n)}},$$

which is equal to the weighted mean of all the samples, weighted according to their class membership probabilities. Update equations for the covariance matrix are derived in a similar fashion as the mean vectors. The resulting update equation is:

$$\mathbf{\Sigma}_c^{(n+1)} = \frac{\sum_{i=1}^{N} p_{ci}^{(n)} (\mathbf{x}_i - \mu_c^{(n+1)})(\mathbf{x}_i - \mu_c^{(n+1)})^T}{\sum_{i=1}^{N} p_{ci}^{(n)}}.$$

5.2.4 EM for Estimating a von-Mises Fisher Mixture Model

The mixture density estimation problem is not limited to Gaussian distributions, but can easily be generalized to many other parametric distributions. In this chapter we are interested in clustering diffusion tensors, which contain directional information. The von-Mises Fisher (vMF) distribution is a simple parametric distribution for directional data, and has properties similar to the Gaussian distribution. For example, the maximum entropy density defined on a unit sphere is a vMF distribution for a fixed mean vector [49].

Similar to the Gaussian mixtures, we define a mixture density estimation problem for the vMF mixtures: Given a set of N points on a D dimensional unit sphere, $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \,|\mathbf{x}_i \in \mathbb{R}^D \text{ and } \|\mathbf{x}_i\| = 1\}$, and a family $\mathscr{G}$ of density probability density functions defined on $\mathbb{R}^D$, calculate the probability density function $g(\mathbf{x}) \in \mathscr{G}$ that max-

imizes the likelihood of the observations in X. For a vMF mixture model, the density functions are conditioned to be mixtures of vMF, i.e.,

$$g(\mathbf{x}) = \sum_{c=1}^{C} \pi_c f(\mathbf{x}|c;\Theta)$$

where π_c is the mixing weight (probability) for mixture c. $f(\mathbf{x}|c;\Theta)$ is the vMF distribution function, i.e.

$$f(\mathbf{x}|c;\Theta) = vMF(\mathbf{x};\nu_c,\kappa_c) = C_D(\kappa_c)exp\left(\kappa_c {\nu_\mathbf{c}}^T \mathbf{x}\right),$$

where ν_c is the mean vector that is subject to the constraint $||\nu_c|| = 1$. κ_c is called the "concentration parameter" for mixture c, since it describes how strongly the unit vectors are drawn toward the mean vector. A higher value of κ_c results in a more concentrated distribution of the vectors on the unit sphere. The normalization constant $C_D(\kappa_c)$ is given by:

$$C_D(\kappa) = \kappa^{(D/2)-1}/(2\pi)^{D/2} I_{(D/2)-1}(\kappa),$$

where $I_r(.)$ represents the modified Bessel function of the first kind and order r. As before, the mixture weights need to sum up to one, $\sum_c \pi_c = 1$ and be non-negative, $\pi_c \geq 0$. Given this model, the log-likelihood of X becomes:

$$\mathscr{L}(\Theta) = \sum_{i=1}^{N} \log \sum_{c=1}^{C} \pi_c vMF(\mathbf{x}_i;\nu_c,\kappa_c).$$

Similar to the Gaussian mixture case, we first obtain a lower bound on the log-likelihood and then derive the specific update equations.

E-Step

In the E-Step the algorithm constructs a lower bound $b_n(\Theta)$ by taking the expectation of the log-likelihood with respect to the current distribution of the class labels. From Equation

5.6, the lower bound (excluding the constant term) is:

$$\begin{aligned}
\widetilde{b}_n(\Theta) &= \sum_{i=1}^{N}\sum_{c=1}^{C} p_{ci}^{(n)} \log p(\mathbf{x}_i, c; \Theta), \\
&= \sum_{i=1}^{N}\sum_{c=1}^{C} p_{ci}^{(n)} \log \pi_c vMF(\mathbf{x}_i; \nu_c, \kappa_c), \\
&= \sum_{i=1}^{N}\sum_{c=1}^{C} p_{ci}^{(n)} \log \pi_c + \sum_{i=1}^{N}\sum_{c=1}^{C} p_{ci}^{(n)} C_D(\kappa_c) + \sum_{i=1}^{N}\sum_{c=1}^{C} p_{ci}^{(n)} \kappa_c \nu_{\mathbf{c}}{}^T \mathbf{x}_i, \quad (5.10)
\end{aligned}$$

where $p_{ci}^{(n)} = p(c|\mathbf{x}_i; \Theta^{(n)})$. To construct this bound, we need to calculate the membership probabilities $p_{ci}^{(n)}$ for each of the samples in X:

$$\begin{aligned}
p_{ci}^{(n)} &= \frac{p(\mathbf{x}_i|c; \Theta^{(n)}) p(c; \Theta^{(n)})}{p(\mathbf{x}_i; \Theta^{(n)})}, \\
&= \frac{\pi_c^{(n)} vMF(\mathbf{x}_i; \nu_c^{(n)}, \kappa_c^{(n)})}{\sum_{c=1}^{C} \pi_c^{(n)} vMF(\mathbf{x}_i; \nu_c^{(n)}, \kappa_c^{(n)})}.
\end{aligned}$$

M-Step

In the M-Step, the algorithm updates the parameter set Θ to maximize the lower bound $b_n(\Theta)$. The update equations for each of the parameters are derived using Lagrange multipliers for the corresponding constraints (if any) and setting the derivative of the lower bound to zero. We first derive the update equation for π_c with the constraint $\sum_c \pi_c = 1$:

$$H_\pi(\Theta) = \widetilde{b}_n(\Theta) - \lambda \left(\sum_{c=1}^{C} \pi_c - 1 \right),$$

Taking the derivative of $H_\pi(\Theta)$ with respect to a specific mixture weight, π_j for $c = j$ results in:

$$\frac{\partial H_\pi(\Theta)}{\partial \pi_j} = \sum_{i=1}^{N} \frac{p_{ji}^{(n)}}{\pi_j} - \lambda.$$

This is the same as Equation 5.9 from the Gaussian mixture model derivation, and results in the same update equation for π_c:

$$\pi_c^{(n+1)} = \frac{1}{N}\sum_{i=1}^{N} p_{ci}^{(n)}.$$

The mean vectors ν_c are constrained to be on the unit sphere, i.e. $\nu_c^T\nu_c = 1$. Note that this constraint is dependent on the class index c, resulting in C equations with C Lagrange multipliers λ_c. Using a Lagrange multiplier for a specific constraint for $c = j$:

$$H_\nu(\Theta) = p_{ci}^{(n)} \log \pi_j + \sum_{i=1}^{N} p_{ji}^{(n)} C_D(\kappa_j) + \sum_{i=1}^{N} p_{ji}^{(n)} \kappa_j \nu_j{}^T \mathbf{x}_i - \lambda_j \left(\nu_j^T \nu_j - 1\right),$$

Taking the derivative of $H_\nu(\Theta)$ with respect to the mean vector, ν_j results in:

$$\frac{\partial H_\nu(\Theta)}{\partial \nu_j} = \sum_{i=1}^{N} p_{ji}^{(n)} \kappa_j \mathbf{x}_i - \lambda_j 2\nu_j.$$

Setting the derivative to zero and rearranging to solve for ν_j gives us:

$$\begin{aligned} \nu_j &= \frac{1}{2\lambda_j}\sum_{i=1}^{N} p_{ji}^{(n)} \kappa_j \mathbf{x}_i, \\ &= \frac{\mathbf{r}_j}{2\lambda_j}, \end{aligned} \tag{5.11}$$

where $\mathbf{r}_j = \sum_i p_{ji}^{(n)} \mathbf{x}_i$, the weighted mean vector. Adding the constraint $\nu_j^T\nu_j = 1$ to solve for λ_j results in:

$$\frac{\kappa_j^2}{4\lambda_j^2}\mathbf{r}_j^T\mathbf{r}_j = 1.$$

Rearranging the equation gives us:

$$\lambda_j = \frac{\kappa_j \|\mathbf{r}_j\|}{2}.$$

Finally, plugging λ_j back into Equation 5.11 gives us the update equation for ν_c:

$$\nu_c^{(n+1)} = \frac{\mathbf{r}_c}{\|\mathbf{r}_c\|} = \frac{\sum_{i=1}^{N} p_{ci}^{(n)} \mathbf{x}_i}{\| \sum_{i=1}^{N} p_{ci}^{(n)} \mathbf{x}_i \|}.$$

The update equation for the concentration parameter κ_c is derived in a similar fashion using the constraint $\kappa_c \geq 0$. The resulting equation is:

$$\frac{I_{(D/2)}(\kappa_c^{(n+1)})}{I_{(D/2)-1}(\kappa_c^{(n+1)})} = \frac{\|\mathbf{r_c}\|}{\sum_{i=1} p_{ci}^{(n)}} = \bar{r}_c.$$

Since the equation contains a fraction of two Bessel function, there is no analytical solution for $\kappa_c^{(n+1)}$. We can either use a numerical scheme to solve for $\kappa_c^{(n+1)}$, or use an approximation. We use an asymptotic approximation proposed in [3]:

$$\kappa_c^{(n+1)} \approx \frac{(D\bar{r}_c - \bar{r}_c^3)}{(1 - \bar{r}_c^2)}.$$

Generalized Expectation Maximization

Even when there is no analytical solution for the maximum of the lower bound, like in the case of κ_c, we can derive an iterative algorithm that shares the same properties with EM, such as the guarantee for convergence to the local minima, as long as the lower bound is increased (but not necessarily maximized) in the M-Step [51]. The resulting algorithm is called "generalized Expectation Maximization" (gEM). However, gEM may require more iterations to converge, since the lower bound is increased but not maximized in the M-Step. In general, the M-Step of gEM is slower to compute, if a numerical scheme is used rather than an analytical solution for the maximization.

5.3 Consistency Clustering

In this section we formulate the problem of population-wise consistent segmentation of thalamic nuclei and derive a solution using gEM as described in Section 5.2. We ensure correspondence of class labels and consistency between subjects through a single mixture model (atlas) for all subjects. Since there is significant variability between the location, orientation, shape and other properties of the nuclei between subjects (See Figure 4-2), we also introduce registration parameters in our framework for spatial normalization. Since the principal diffusion direction results in more accurate segmentations than full tensor based metrics (Chapter 4), we use only the principal diffusion direction and the spatial location of the tensors as features in the segmentation.

We define a maximum likelihood density estimation problem as following: Given a set of N diffusion tensors from S subjects, where each tensor is represented by spatial location $X = \{X_s\}$, $X_s = \{\mathbf{x}_i | \mathbf{x}_i \in \mathbb{R}^3\}$ and principal diffusion direction $V = \{V_s\}$, $V_s = \{\mathbf{v}_i | \mathbf{v}_i \in \mathbb{R}^3 \text{ and } \|\mathbf{v}_i\| = 1\}$, and a family $\mathscr{G}$ of density probability density functions defined on $\mathbb{R}^3 \cup \mathbb{R}^3$, calculate the joint probability density function $p(\mathbf{x}, \mathbf{v}; \Theta) \in \mathscr{G}$ that maximizes the likelihood of the observations in $\{X, V\}$. We model the DTI data with the following set of parameters: $\Theta = \{\pi_c, \mu_c, \mathbf{\Sigma}_c, \nu_c, \kappa_c\} \cup \{\mathbf{R}_s\}$, where c is an index over class labels, i.e. $c \in \{1, 2, \ldots, C\}$ and s is an index over subjects, i.e. $s \in \{1, 2, \ldots, S\}$. Given this model, our problem is finding the optimal parameter set Θ^* that maximizes the likelihood of observing the sample pairs $\{X, V\}$. Given these parameters, the log-likelihood of the parameter set becomes:

$$\mathscr{L}(\Theta) = \sum_{s=1}^{S} \sum_{\mathbf{x}_i \in X_s} \log \sum_{c=1}^{C} \pi_c f_x(\mathbf{R}_s \circ \mathbf{x}_i; \Theta) f_v(\mathbf{R}_s \circ \mathbf{v}_i; \Theta),$$

where we assume independence between every observed sample pair $\{\mathbf{x}_i, \mathbf{v}_i\}$ and also independence between the spatial location $\mathbf{x}_i$ and principal diffusion direction $\mathbf{v}_i$. $\mathbf{R}_s \circ \mathbf{x}_i$ and $\mathbf{R}_s \circ \mathbf{v}_i$ denote the sample pairs after undergoing the spatial registration for subject s. To derive the specific equations for this problem, we also need to choose a family of parametric

density functions. We model the spatial locations with Gaussian density functions:

$$f_x(\mathbf{x};\Theta) = f_x(\mathbf{x};\mu_c,\mathbf{\Sigma}_c) = \frac{1}{(2\pi)^{3/2}|\mathbf{\Sigma}_c|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_c)^T\mathbf{\Sigma}_c^{-1}(\mathbf{x}-\mu_c)\right),$$

where μ_c is the mean vector and $\mathbf{\Sigma}_c$ is the covariance matrix. We model the distribution of the principal diffusion directions with a von Mises-Fisher distribution:

$$f_v(\mathbf{v};\Theta) = f_v(\mathbf{v};\nu_c,\kappa_c) = C(\kappa_c)exp\left(\kappa_c{\nu_\mathbf{c}}^T\mathbf{v}\right),$$

where ν_c is the mean orientation and κ_c is the concentration parameter. The constant, $C(\kappa) = \kappa^{1/2}/(2\pi)^{3/2}I_{1/2}(\kappa)$, where $I_{1/2}(\kappa)$ is a modified Bessel function of the first kind and order $1/2$. Under this model, we formulate our problem as a maximum likelihood estimation of the parameter set Θ:

$$\Theta^* = \arg\max_{\Theta} \mathscr{L}(\Theta).$$

In the next sections we obtain a lower bound for the log likelihood and derive specific update equations for our formulation to iteratively estimate Θ^*.

5.3.1 E-Step

In the E-Step the CC constructs the lower bound $b_n(\Theta)$ by taking the expectation of the log-likelihood with respect to the current distribution of the class labels. In this case, the

lower bound (excluding the constant term) is:

$$\begin{aligned}
\widetilde{b}_n(\Theta) &= \sum_{s=1}^{S}\sum_{\mathbf{x}_i\in X_s}\sum_{c=1}^{C} p_{ci}^{(n)} \log p(\mathbf{x}_i, \mathbf{v}_i, c; \Theta), \\
&= \sum_{s=1}^{S}\sum_{\mathbf{x}_i\in X_s}\sum_{c=1}^{C} p_{ci}^{(n)} \log \pi_c f_x(\mathbf{R}_s \circ \mathbf{x}_i; \mu_c, \mathbf{\Sigma}_c) f_v(\mathbf{R}_s \circ \mathbf{v}_i; \nu_c, \kappa_c), \\
&= \sum_{s=1}^{S}\sum_{\mathbf{x}_i\in X_s}\sum_{c=1}^{C} p_{ci}^{(n)} \log \pi_c + \sum_{s=1}^{S}\sum_{\mathbf{x}_i\in X_s}\sum_{c=1}^{C} p_{ci}^{(n)} \log f_x(\mathbf{R}_s \circ \mathbf{x}_i; \mu_c, \mathbf{\Sigma}_c) \\
&\quad + \sum_{s=1}^{S}\sum_{\mathbf{x}_i\in X_s}\sum_{c=1}^{C} p_{ci}^{(n)} \log f_v(\mathbf{R}_s \circ \mathbf{v}_i; \nu_c, \kappa_c), \qquad (5.12)
\end{aligned}$$

where $p_{ci}^{(n)} = p(c|\mathbf{x}_i, \mathbf{v}_i; \Theta^{(n)})$. To construct this bound, we need to calculate the membership probabilities $p_{ci}^{(n)}$ for each of the samples pairs in $\{X, V\}$:

$$\begin{aligned}
p_{ci}^{(n)} &= \frac{p(\mathbf{x}_i|c; \Theta^{(n)}) p(c; \Theta^{(n)})}{p(\mathbf{x}_i; \Theta^{(n)})}, \\
&= \frac{\pi_c^{(n)} f_x(\mathbf{R}_s^{(n)} \circ \mathbf{x}_i; \mu_c^{(n)}, \mathbf{\Sigma}_c^{(n)}) f_v(\mathbf{R}_s^{(n)} \circ \mathbf{v}_i; \nu_c^{(n)}, \kappa_c^{(n)})}{\sum_{c=1}^{C} \pi_c^{(n)} f_x(\mathbf{R}_s^{(n)} \circ \mathbf{x}_i; \mu_c^{(n)}, \mathbf{\Sigma}_c^{(n)}) f_v(\mathbf{R}_s^{(n)} \circ \mathbf{v}_i; \nu_c^{(n)}, \kappa_c^{(n)})}. \qquad (5.13)
\end{aligned}$$

5.3.2 M-Step

In the M-Step, the CC updates the parameter set Θ to maximize the lower bound $b_n(\Theta)$. The update equations for each of the parameters are derived using Lagrange multipliers for the corresponding constraints (if any) and setting the derivative of the lower bound to zero. We note that the lower bound of Equation 5.12 is similar to the lower bounds in the Gaussian (Equation 5.7) and vMF mixture (Equation 5.10) cases. We rewrite (after change of variables) the lower bound of CC in a form that is similar to the previous mixture models, which result in similar update equations as in Sections 5.2.3 and 5.2.4:

$$\begin{aligned}
\widetilde{b}_n(\Theta) &= \sum_{i=1}^{N}\sum_{c=1}^{C} p_{ci}^{(n)} \log \pi_c + \sum_{i=1}^{N}\sum_{c=1}^{C} p_{ci}^{(n)} \log f_x(\mathbf{R} \circ \mathbf{x}_i; \mu_c, \mathbf{\Sigma}_c) \\
&\quad + \sum_{i=1}^{N}\sum_{c=1}^{C} p_{ci}^{(n)} \log f_v(\mathbf{R} \circ \mathbf{v}_i; \nu_c, \kappa_c). \qquad (5.14)
\end{aligned}$$

We note that π_c appears only in the first term of the lower bound, and is in same format as in the previous cases. Therefore, the update equation for $\pi_c^{(n+1)}$ is the same as before:

$$\begin{aligned} \pi_c^{(n+1)} &= \frac{1}{N}\sum_{i=1}^{N} p_{ci}^{(n)}, \\ &= P_c^{(n)}/N, \end{aligned} \tag{5.15}$$

where we define $P_c^{(n)} = \sum_{i=1}^{N} p_{ci}^{(n)}$ for a more compact notation. Similarly, the Gaussian mixture parameters μ_c and Σ_c appear only in the second term of the lower bound of Equation 5.14, and are in the same format as in the Gaussian mixture model case. Therefore, the update equations for these two parameters are:

$$\mu_c^{(n+1)} = \frac{1}{P_c^{(n)}}\sum_{i=1}^{N} p_{ci}^{(n)}\mathbf{x}_i, \tag{5.16}$$

$$\Sigma_c^{(n+1)} = \frac{1}{P_c^{(n)}}\sum_{i=1}^{N} p_{ci}^{(n)}\left(\mathbf{x}_i - \mu_c^{(n+1)}\right)\left(\mathbf{x}_i - \mu_c^{(n+1)}\right)^T. \tag{5.17}$$

$$\tag{5.18}$$

The vMF mixture parameters only appear in the last term of the lower bound, and the resulting update equations are the same as the ones in the vMF mixture model case:

$$\mathbf{r}_c = \sum_{i=1}^{N} p_{ci}^{(n)}\mathbf{v}_i, \tag{5.19}$$

$$\bar{r}_c = \frac{\|\mathbf{r}_c\|}{P_c^{(n)}}, \tag{5.20}$$

$$\nu_c^{(n+1)} = \frac{\mathbf{r}_c}{\|\mathbf{r}_c\|}, \tag{5.21}$$

$$\kappa_c^{(n+1)} \approx \frac{(3\bar{r}_c - \bar{r}_c^3)}{(1-\bar{r}_c^2)}, \tag{5.22}$$

where, in the last equation, we plugged in $D = 3$ since our problem is three dimensional.

Registration

Registration parameters are also updated in the M-Step. We parametrize the registration as one rigid transformation per class per subject, i.e.,

$$\begin{aligned}\mathbf{R}_s^{(n)} \circ \mathbf{x}_i &= \mathbf{R}_{sc}^{(n)}(\mathbf{x}_i - \mu_{sc}^{(n)}) + \mu_{sc}^{(n)} + \mathbf{t}_{sc}^{(n)}, \\ \mathbf{R}_s^{(n)} \circ \mathbf{v}_i &= \mathbf{R}_{sc}^{(n)}\mathbf{v}_i,\end{aligned}$$

where $\mathbf{R}_{sc}^{(n)}$ is a rotation matrix and $\mu_{sc}^{(n)}$ is the weighted mean of the voxel locations in a given subject; i.e., $\mu_{sc}^{(n)} = \sum_{i \in s} p_{ci}^{(n)} \mathbf{x}_i / \sum_{i \in s} p_{ci}^{(n)}$. We first derive the update equation for the translation components of the registration. Rewriting the items that depend on the translation $\mathbf{t}_{sc}$ from the lower bound of Equation 5.12, we get:

$$H_{\mathbf{t}}(\mathbf{t}_{sc}) = \sum_{s=1}^{S} \sum_{\mathbf{x}_i \in X_s} \sum_{c=1}^{C} p_{ci}^{(n)} \log f_x(\mathbf{R}_s \circ \mathbf{x}_i; \mu_c, \mathbf{\Sigma}_c).$$

Fixing the subject index $s = k$ and the class index $c = j$ for a particular translation component $\mathbf{t}_{kj}$, we obtain:

$$H_{\mathbf{t}}(\mathbf{t}_{kj}) = \sum_{\mathbf{x}_i \in X_k} \left(-\frac{1}{2}(\mathbf{R}_k \circ \mathbf{x}_i - \mu_j^{(n+1)})^T (\mathbf{\Sigma}_j^{(n+1)})^{-1} (\mathbf{R}_k \circ \mathbf{x}_i - \mu_j^{(n+1)}) + C(f_{\mathbf{x}}) \right),$$

where we plugged in the most current estimates for μ_j and $\mathbf{\Sigma}_j$. $C(f_{\mathbf{x}})$ is the constant term that does not depend on $\mathbf{t}_{kj}$. We recall that $\mathbf{R}_k \circ \mathbf{x}_i = \mathbf{R}_{kj}(\mathbf{x}_i - \mu_{kj}) + \mu_{kj} + \mathbf{t}_{kj}$. Let $\bar{\mathbf{x}}_i = \mathbf{x}_i - \mu_{kj}$ and $\bar{\mathbf{t}}_{kj} = \mu_{kj} + \mathbf{t}_{kj}$, then ignoring the constant term, our problem becomes:

$$H_{\mathbf{t}}(\mathbf{t}_{kj}) = \sum_{l=1}^{N_k} p_{jl}^{(n)} \left((\mathbf{R}_{kj}\bar{\mathbf{x}}_l + \bar{\mathbf{t}}_{kj} - \mu_j^{(n+1)})^T (\mathbf{\Sigma}_j^{(n+1)})^{-1} (\mathbf{R}_{kj}\bar{\mathbf{x}}_l + \bar{\mathbf{t}}_{kj} - \mu_j^{(n+1)}) \right),$$

where N_k is the number of samples in subject k, and l is a replacement index for notational convenience. Taking the derivative of $H_{\mathbf{t}}(\mathbf{t}_{kj})$ with respect to $\bar{\mathbf{t}}_{kj}$ yields:

$$\frac{\partial H_{\mathbf{t}}(\mathbf{t}_{kj})}{\partial \bar{\mathbf{t}}_{kj}} = 2(\mathbf{\Sigma}_j^{(n+1)})^{-1} \sum_{l=1}^{N_k} p_{jl}^{(n)} \left(\mathbf{R}_{kj}\bar{\mathbf{x}}_l + \bar{\mathbf{t}}_{kj} - \mu_j^{(n+1)} \right).$$

We set the derivative to zero and rearrange the equation to solve for $\bar{\mathbf{t}}_{kj}$:

$$\bar{\mathbf{t}}_{kj} \sum_{l=1}^{N_k} p_{jl}^{(n)} = \mu_j^{(n+1)} \sum_{l=1}^{N_k} p_{jl}^{(n)} - \mathbf{R}_{kj} \left(\sum_{l=1}^{N_k} p_{jl}^{(n)} \bar{\mathbf{x}}_l \right).$$

Since $\sum_{l=1}^{N_k} p_{jl}^{(n)} \bar{\mathbf{x}}_l = 0$;

$$\bar{\mathbf{t}}_{kj} = \mu_j^{(n+1)}.$$

Then, the update equation for the translation component $\mathbf{t}_{sc}^{(n+1)}$ becomes:

$$\mathbf{t}_{sc}^{(n+1)} = \mu_c^{(n+1)} - \mu_{sc}^{(n+1)}.$$

Thus, the optimal translation is the one that aligns the weighted mean of the voxels in subject s and the weighted mean of the voxels in all subjects. Unfortunately, the same technique does not lead to a simple analytical solution for the rotation matrices, $\mathbf{R}_{sc}^{(n+1)}$. However, we derive a maximum likelihood optimization function and optimize the function using a numerical scheme. The resulting optimization function is:

$$\begin{aligned} \mathbf{R}_{sc}^{(n+1)} &= \arg\max_{\mathbf{R}_{sc}} \sum_{\mathbf{x}_i \in X_s} p_{ci}^{(n)} \left(2\kappa_c^{(n+1)} \nu_c^{(n+1),T} \mathbf{R}_{sc} \mathbf{v}_i - \mathbf{x}_i^T \mathbf{R}_{sc}^T \mathbf{\Sigma}_c^{(n+1),T} \mathbf{R}_{sc} \mathbf{x}_i \right) \\ &\text{s.t.} \quad \mathbf{R}_{sc} \mathbf{R}_{sc}^T = I \text{ and } |R| = 1. \end{aligned}$$

We further parametrize $\mathbf{R}_{sc}$ using Euler angles so that the constraints are automatically met. Then we find optimal values for the Euler angles (and therefore $\mathbf{R}_{sc}$) using a simplex search method [40].

5.3.3 Prior Information

Our model is formulated as a maximum likelihood problem and is well suited for population-wise clustering without any labeled data. However, in practice, it may be possible to obtain a few labeled subjects or just a few labeled voxels in one subject without much additional cost. In that case, it is interesting to use this prior information in the segmentation. To enable the use of prior information, we modify our problem, and show that the resulting

algorithm is a simple variation of the algorithm derived above.

The density estimation problem in the presence of labeled data becomes: Given a set of N diffusion tensors from S subjects, where each tensor is represented by spatial location $X = \{X_s\}$, $X_s = \{\mathbf{x}_u | \mathbf{x}_u \in \mathbb{R}^3\} \cup \{\mathbf{x}_l | \mathbf{x}_l \in \mathbb{R}^3\}$, where $\mathbf{x}_u$ denote the unlabeled data and $\mathbf{x}_l$ denote the labeled data. Also given a set of principal diffusion direction $V = \{V_s\}$, $V_s = \{\mathbf{v}_i | \mathbf{v}_i \in \mathbb{R}^3 \text{ and } \|\mathbf{v}_i\| = 1\}$, where $\mathbf{v}_u$ denote the unlabeled data and $\mathbf{v}_l$ denote the labeled data. Calculate the joint probability density function $p(\mathbf{x}, \mathbf{v}; \Theta) \in \mathscr{G}$ that maximizes the likelihood of the observations in $\{X, V\}$. In this case, the log-likelihood of the parameter set becomes:

$$\begin{aligned}
\mathscr{L}(\Theta) &= \alpha \sum_{s=1}^{S} \sum_{\mathbf{x}_l \in X_s} \log \sum_{c=1}^{C} \pi_c f_x(\mathbf{R}_s \circ \mathbf{x}_l; \Theta) f_v(\mathbf{R}_s \circ \mathbf{v}_l; \Theta) \\
&\quad + (1-\alpha) \sum_{s=1}^{S} \sum_{\mathbf{x}_u \in X_s} \log \sum_{c=1}^{C} \pi_c f_x(\mathbf{R}_s \circ \mathbf{x}_u; \Theta) f_v(\mathbf{R}_s \circ \mathbf{v}_u; \Theta),
\end{aligned}$$

where we introduced a weighting parameter α that controls the impact of labeled data on the learned model. When $\alpha = 0$, the labeled data is not used by the algorithm and this formulation becomes exactly the same as before. When $\alpha = 0.5$, the labeled and unlabeled data are weighted equally, and labeled data is used to "anchor" the model. When $\alpha = 1$, the model is built solely by the labeled data. We set $\alpha = 0.5$ in our experiments, unless it is stated otherwise.

Rewriting the lower bound from Equation 5.14 to reflect this new formulation results:

$$\begin{aligned}
\widetilde{b}_n(\Theta) &= \alpha \sum_{l=1}^{N} \sum_{c=1}^{C} p_{cl} \left(\log \pi_c + \log f_x(\mathbf{R} \circ \mathbf{x}_l; \mu_c, \mathbf{\Sigma}_c) + \log f_v(\mathbf{R} \circ \mathbf{v}_l; \nu_c \kappa_c) \right) \\
&\quad + (1-\alpha) \sum_{u=1}^{N} \sum_{c=1}^{C} p_{cu}^{(n)} \left(\log \pi_c + \log f_x(\mathbf{R} \circ \mathbf{x}_u; \mu_c, \mathbf{\Sigma}_c) + \log f_v(\mathbf{R} \circ \mathbf{v}_u; \nu_c \kappa_c) \right),
\end{aligned}$$

where $p_{cl} = p(c|\mathbf{x}_l, \mathbf{v}_l)$ does not depend on the iteration (n), since the membership probabilities of labeled data are held constant through the model learning. Note that p_{cl} is a probabilistic vector and can be used for both soft and hard priors. When we set $\alpha = 0.5$, the E-Step becomes exactly the same as before; however, the membership probabilities are

updated only for the unlabeled data. The M-Step is also the same as before, with the model updated using both labeled and unlabeled data pooled together.

For $\alpha \neq 0.5$, the E-Step of the algorithm always remains the same; however, the membership probabilities of the labeled data are not updated at any time. The equations for the M-Step also remain the same, but it is easy to see from the lower bound (by moving α and $(1-\alpha)$ inside the summations) that the membership probabilities will need to be weighted by α for the labeled data and $(1-\alpha)$ for the unlabeled data. Therefore, the Consistency Clustering framework can easily handle labeled and unlabeled data together within the same framework, resulting in a semi-supervised technique for atlas creation in diffusion MRI.

5.3.4 Implementation Details

There are certain details that are important in the application of the CC to the DTI data. The first one is about handling the principal diffusion direction, whose sign is arbitrary since the Gaussian probability model of the diffusion tensor is symmetrical around the origin. The second point is about topology correction after convergence of the algorithm.

Handling Ambiguity in the Sign of Principal Diffusion Direction

The Gaussian probability model of the diffusion tensor is symmetrical around the origin, resulting in equal probability of diffusion in direction $\mathbf{v}$ and in direction $-\mathbf{v}$. This results in an arbitrary sign for the principal diffusion direction, and it depends on the specific eigenvector solver used. This is not a problem for a single diffusion tensor; however, since we learn the parameters of the vMF distribution from a multitude of samples, the probability model we learn would become arbitrary if the vector signs are not handled properly.

We formulate the sign ambiguity problem as following: Given a set of N points on a D dimensional unit sphere, $V = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N \,|\mathbf{v}_i \in \mathbb{R}^D \text{ and } \|\mathbf{v}_i\| = 1\}$, and vMF distribution function parameters $\{\nu, \kappa\}$, maximize the log-likelihood of the observations in

V:

$$\arg\max_{s_i} \sum_{i=1}^{N} \log p(s_i \mathbf{v}_i; \nu, \kappa),$$

where $s_i \in \{-1, +1\}$. This formulation leads to calculating the probability of each vector $\mathbf{v}_i$ under the given vMF model with either sign choice, and picking the one that has the higher probability. Since $p(s_i \mathbf{v}_i; \nu, \kappa)$ are independent, we can solve this for each i separately:

$$\begin{aligned}
&\arg\max_{s_i} && \log p(s_i \mathbf{v}_i; \nu, \kappa), \\
&\arg\max_{s_i} && \kappa \nu^T s_i \mathbf{v}_i + \text{constant}, \\
&\arg\max_{s_i} && s_i \nu^T \mathbf{v}_i,
\end{aligned}$$

where we plugged in the specific equation for the log-likelihood and removed the constant term. The last equation is the definition of the modulus function, since $|\nu^T \mathbf{v}_i| \geq \pm \nu^T \mathbf{v}_i$. Therefore, the choice of s_i that maximizes the log-likelihood is:

$$s_i = \text{sgn}(\nu^T \mathbf{v}_i),$$

where $\text{sgn}(.)$ denotes the sign function. Thus, there exists a unique sign for each of the principal diffusion directions that maximizes the log-likelihood and that sign depends on the mean vector of the current vMF model. We handle the sign ambiguity by choosing a unique sign through $s_i = \text{sgn}(\nu^T \mathbf{v}_i)$ when calculating the probability of principal diffusion directions in the E-Step, and similarly in the M-Step, when maximizing the expected log-probability.

Topology Correction

The parametric model we propose is well suited to handle noisy data such as the diffusion tensor images, since the problem of atlas construction is reduced to a parametric estimation problem with few parameters. However, the simplicity of this model may also lead to fragmented segmentations. We handle this through a topology correction step, which is based

on connected components. Topology correction works by identifying the largest spatially connected component for each class, and relabeling any fragments. Fragments are identified as voxels that are not connected to the largest connected component for each class. The relabeling is done using the learned probability model; however, this time assigning to the second most likely class instead of the most likely one, which lead into the fragmentation. In our experiments, we observed that less than 2% of the voxels in the thalamic data are fragments, and can easily be corrected through the topology correction. The model is updated one last time through the M-Step after the topology correction, in order to reflect the changes made to the solution.

5.4 Experiments

The Consistency Clustering was validated on 10 normal subjects' DTI datasets, using only the left hemisphere from each subject. The main questions we were seeking to answer were:

1. Is there an advantage of clustering a population of subjects jointly versus individually?

2. Does prior information increase the accuracy of segmentations in Consistency Clustering?

3. How does Consistency Clustering compare with some of the previously proposed DTI segmentation tools?

5.4.1 Data Collection and Pre-Processing

The same DTI data as in Chapter 4 was used in the experiments. Please refer to Section 4.5 for details on the image acquisition parameters, pre-processing, drawing of the thalamus masks, expert labeling and validation methods.

5.4.2 Individual versus Population-wise Segmentation

The first set of experiments investigated the advantages of population-wise segmentation over individual segmentation. In this first set of experiments, we applied the CC to 10 thalami individually. For each thalamus, the CC was initialized uniformly. The resulting segmentations are presented in Figure 5-3, in which the colors indicate the mean diffusion orientation within each cluster. We also overlaid contours (circles, rectangles and triangles) to indicate inconsistencies between segmentations of different subjects.

The same thalami were also segmented simultaneously using the Consistency Clustering. In this case, the CC was initialized uniformly as in the individual segmentation case. The resulting segmentations are presented in Figure 5-4. The contours of Figure 5-3 were

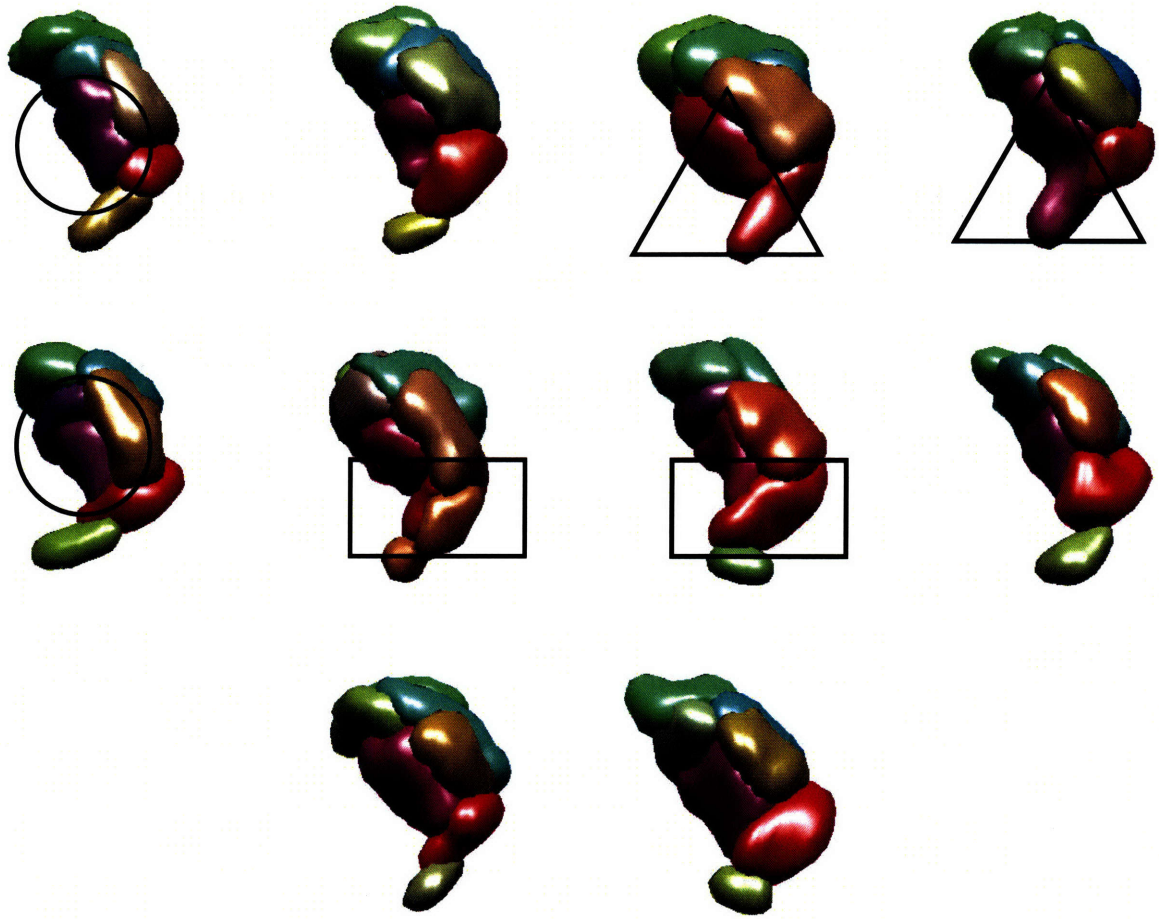

Figure 5-3: Individually segmented thalamic nuclei, resulting from Consistency Clustering with uniform initialization. Each image depicts a segmentation result obtained from a different subject. Colors indicate the mean diffusion orientation within each cluster. Geometric contour pairs indicate areas where segmentations are not consistent among subjects.

once again overlaid in identical locations onto the segmentations. The simultaneous application of the CC was able to correct the inconsistencies in the segmentations. This qualitative improvement in the segmentations also resulted in increased average volume overlaps as well as decreased variance, when compared against the expert labeled data (Figure 5-5).

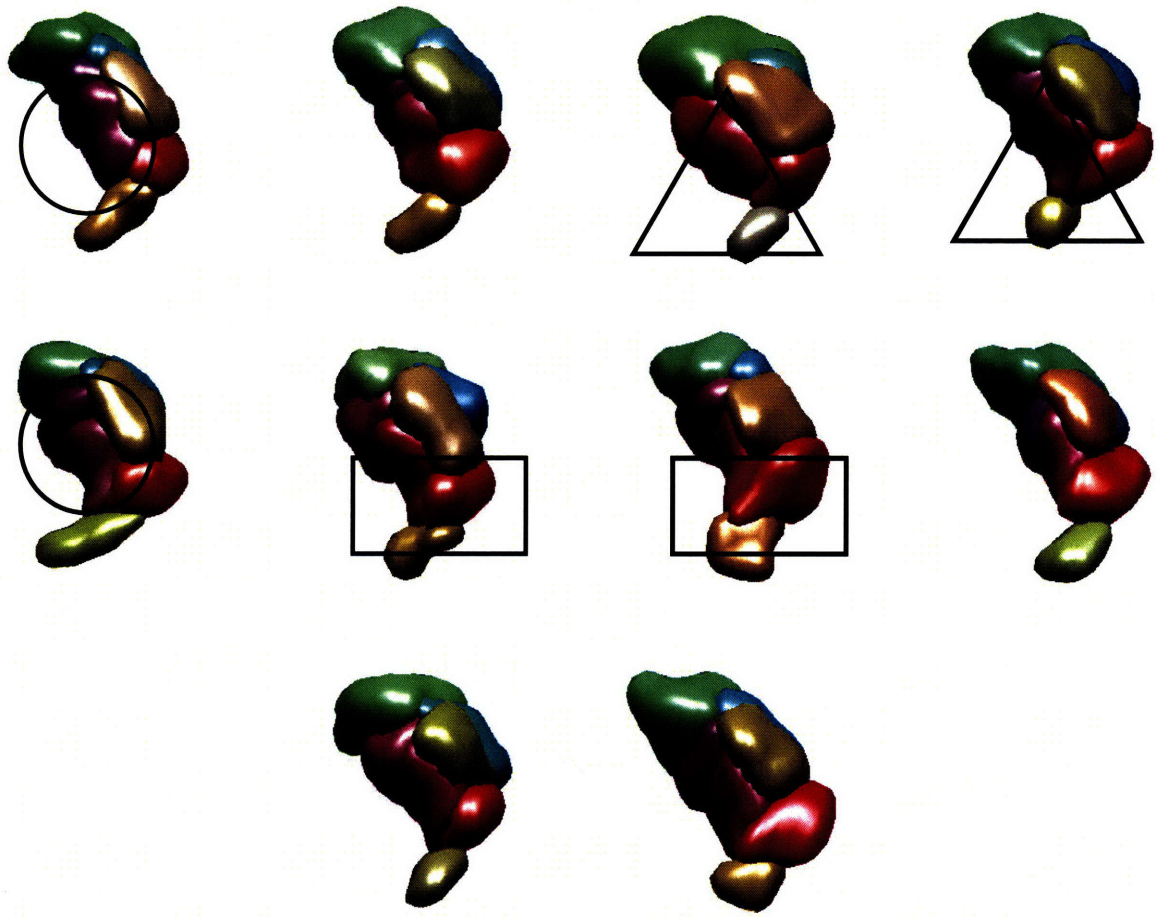

Figure 5-4: Simultaneously segmented thalamic nuclei, resulting from Consistency Clustering with uniform initialization. Each image depicts a segmentation result obtained from a different subject. Colors indicate the mean diffusion orientation within each cluster. Geometric contour pairs are placed in the same locations as in Figure 5-3 and highlight the improvement in consistency of segmentations among subjects

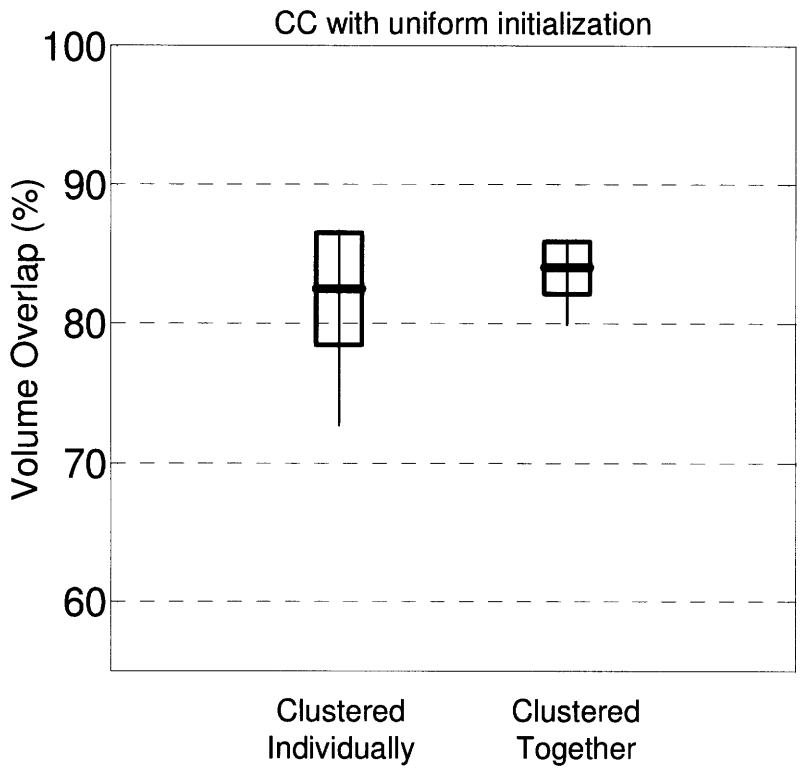

Figure 5-5: Quantitative comparison of individual versus population-wise segmentations obtained through Consistency Clustering with uniform initialization. The boxes indicate one standard deviation around the mean, and the thin lines indicate the range of volume overlaps obtained from 10 subjects. Population-wise segmentation results in increased average volume overlaps as well as decreased variance in volume overlaps.

5.4.3 Effect of Varying Amounts of Prior Information

To understand the effect of varying amounts of prior information on the segmentation quality we repeatedly applied the Consistency Clustering to 10 thalami (simultaneously) using differing amounts of prior information.

The weakest form of prior information was obtained through labeling only k (seven) voxels in one subject, and leaving all remaining voxels unlabeled. These voxels were chosen by first fitting the model of Consistency Clustering to one of the expert labeled subjects, and then identifying the highest probability voxels for each class. The class membership probabilities for each of these voxels were set to be 1 for classes labeled by an expert, and 0 for all other classes (hard priors). These points were then used to initialize the Consistency Clustering, and their membership probabilities were held constant throughout the Consistency Clustering, as described in Section 5.3.3. The resulting segmentation is presented in Figure 5-6.

A stronger form of prior information was obtained by using all the expert labeled voxels in one subject and leaving the other nine subjects unlabeled. The labeled subject was used to initialize the model. The resulting segmentation is presented in Figure 5-7.

To evaluate the impact of the choice of labeled subject on the segmentation results, we repeated this experiment with using different expert-labeled subjects. Since there were ten thalami in our database, there were ten experiments in total. The resulting volume overlaps from each of these experiments are shown as error bars in Figure 5-8. The range and average accuracy of segmentations vary with the choice of expert labeled subject.

The strongest form of prior information was obtained through using nine expert-labeled thalami and labeling the remaining subject through Consistency Clustering. Since applying Consistency Clustering this way would only label one novel subject at a time, this experiment was repeated ten times, once for each novel subject. The resulting segmentation is presented in Figure 5-9. The expert labeled thalami are presented in Figure 5-10 for a qualitative comparison with each of these methods. Consistency clustering was able to segment all subjects with higher accuracy as more prior information was utilized. This observation was also quantified in Figure 5-11 as volume overlaps.

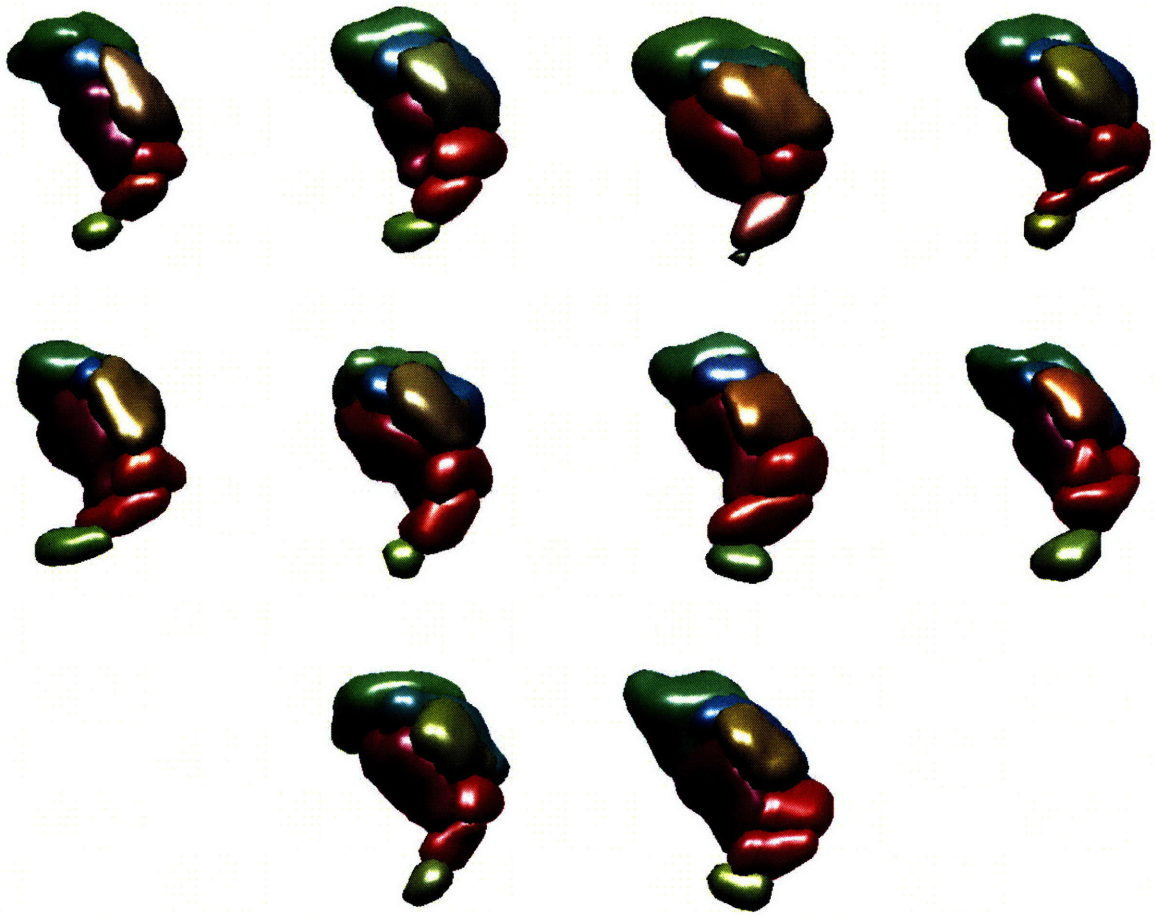

Figure 5-6: Results from Consistency Clustering with seven labeled voxels used as prior information. Each image depicts a segmentation result obtained from a different subject. Colors indicate the mean diffusion orientation within each cluster.

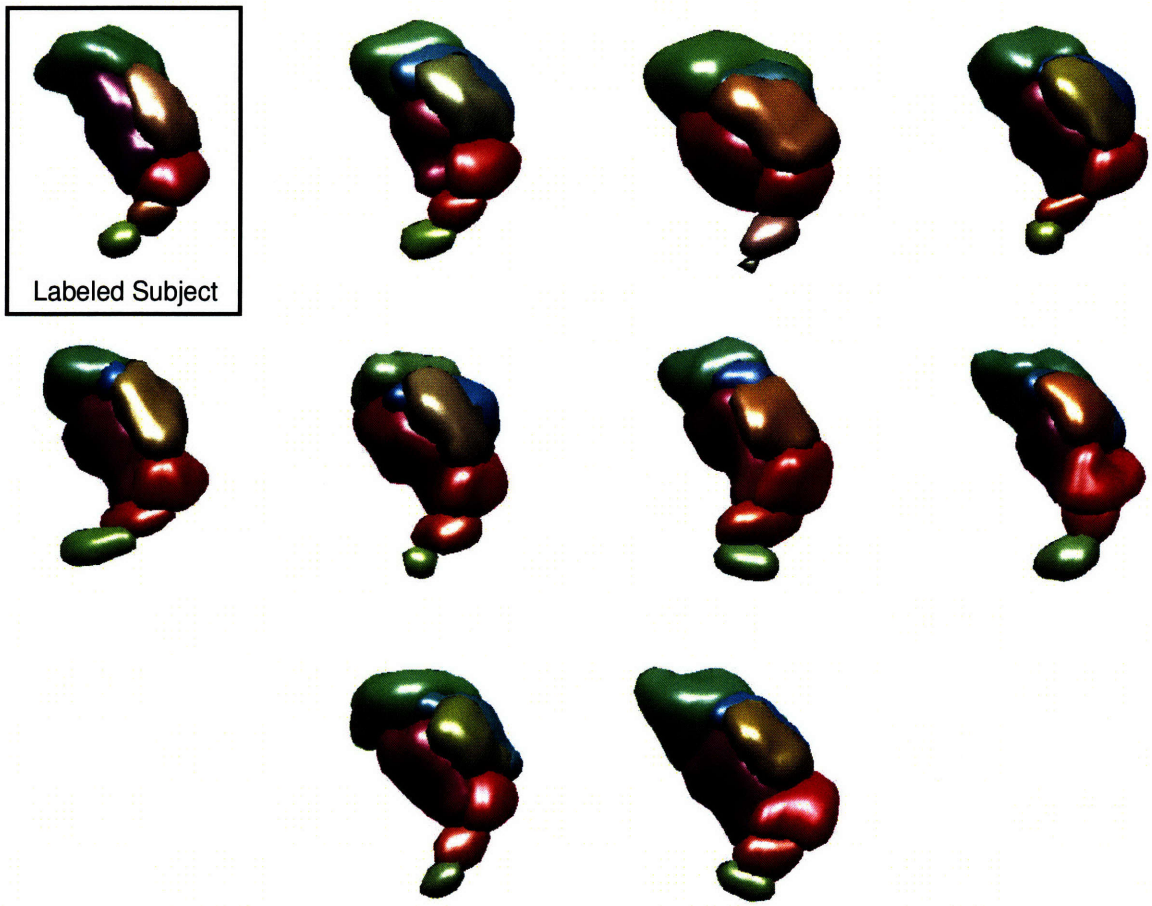

Figure 5-7: Results from Consistency Clustering with single labeled subject used as prior information. Each image depicts a segmentation result obtained from a different subject. Colors indicate the mean diffusion orientation within each cluster.

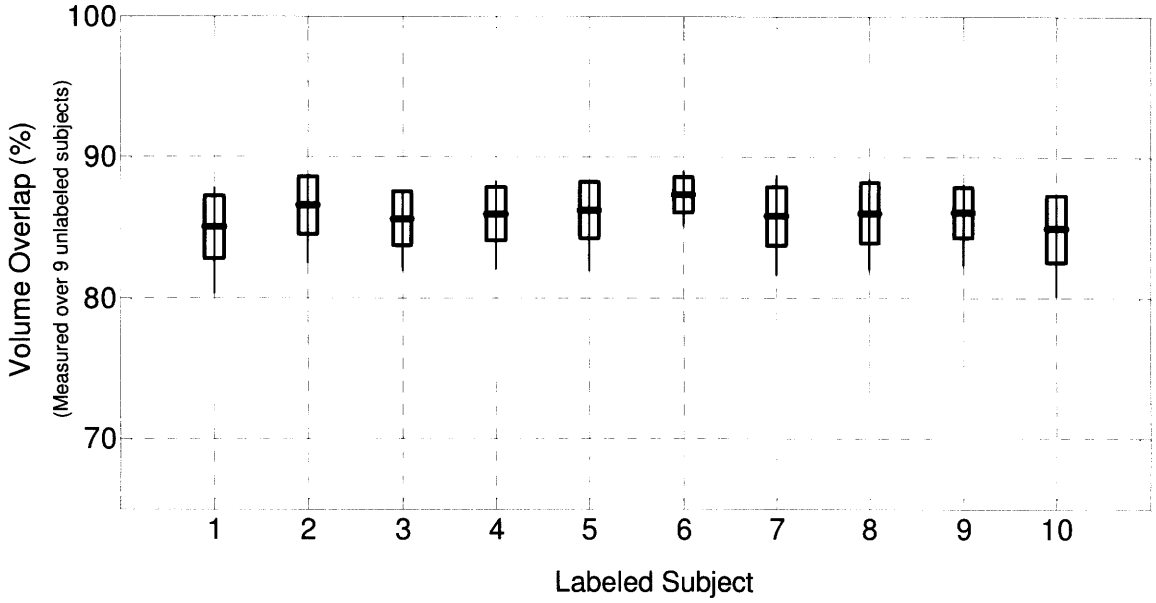

Figure 5-8: Quantitative comparison of differing expert-labeled thalami used as prior information in Consistency Clustering. The boxes indicate one standard deviation around the mean, and the thin lines indicate the range.

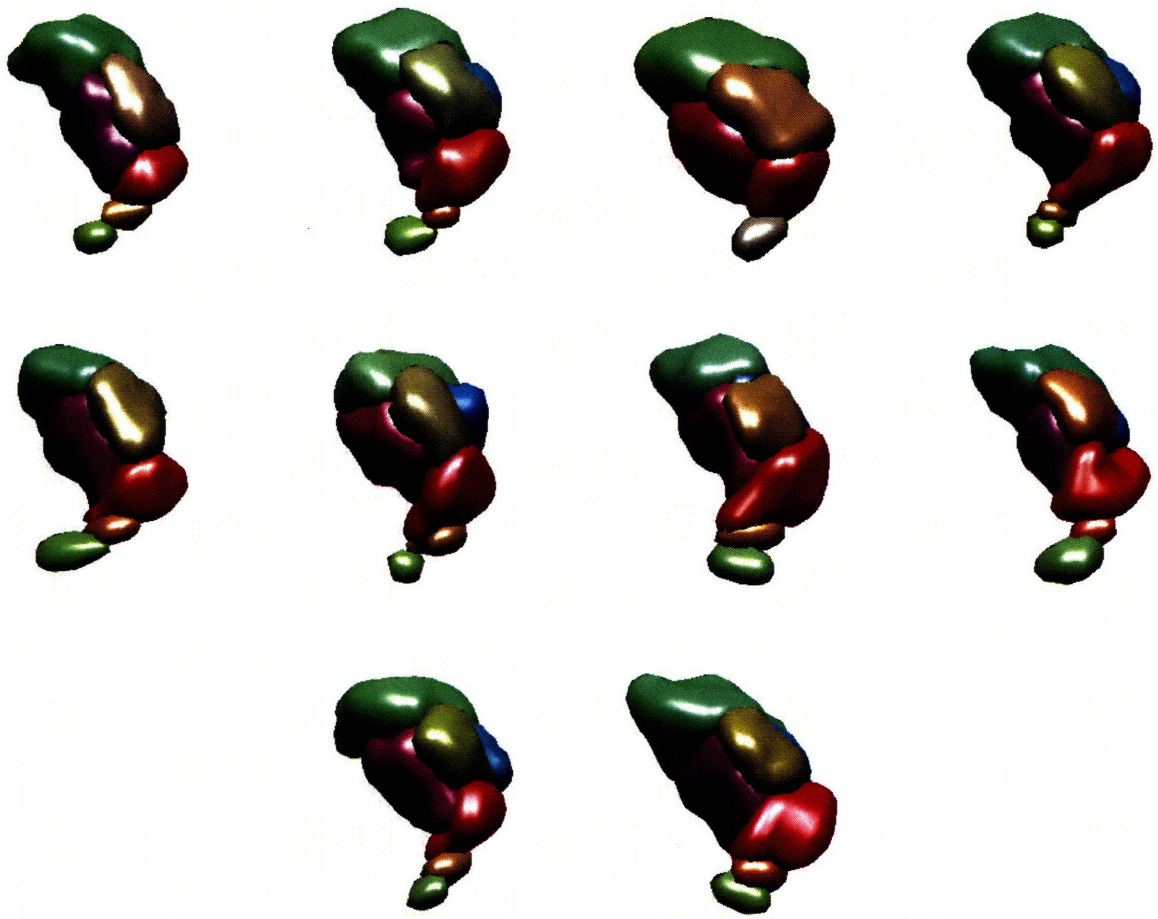

Figure 5-9: Results from Consistency Clustering with nine labeled subjects used as prior information. Segmentation was done in a leave-one-out fashion with nine labeled subjects and one unlabeled subject. The experiment was repeated 10 times, once for each subject shown in the figure. Colors indicate the mean diffusion orientation within each cluster.

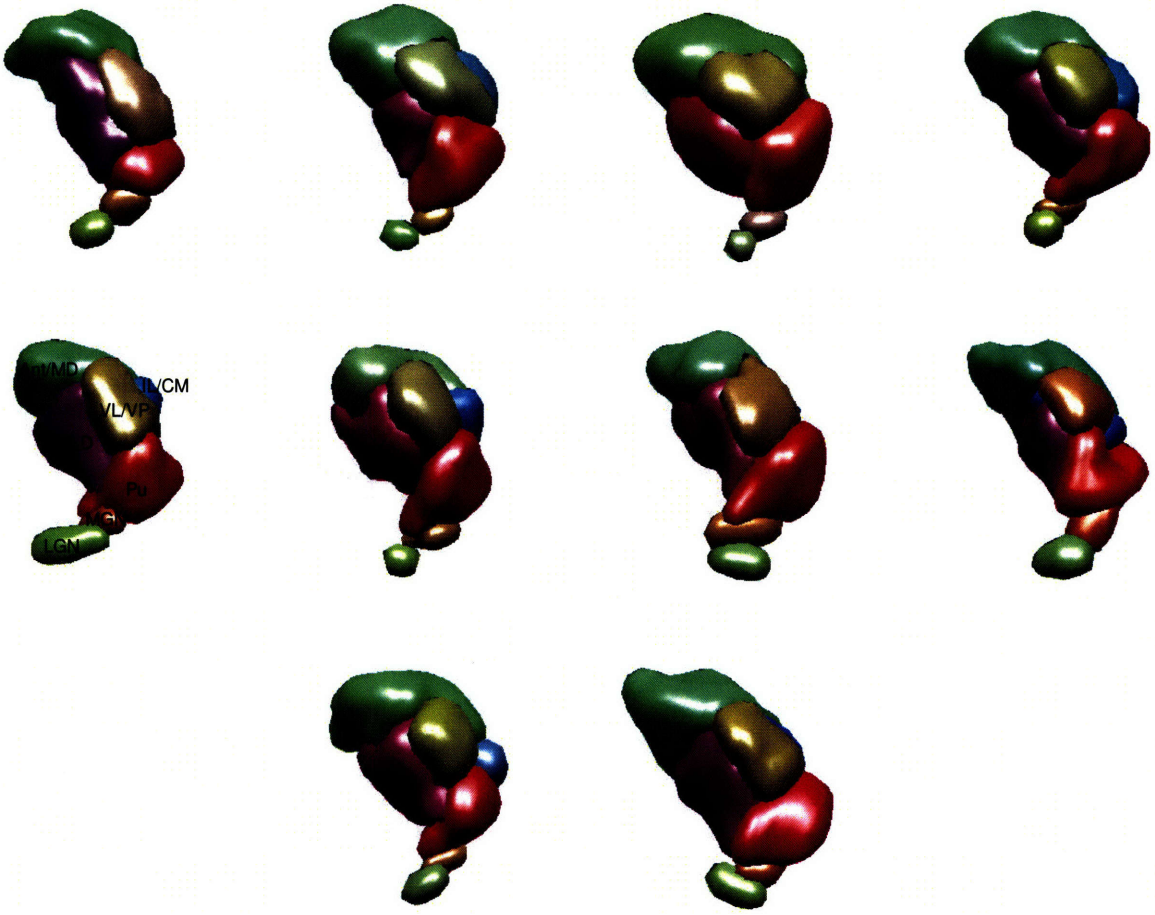

Figure 5-10: Expert labeled thalami. Each image depicts a segmentation result obtained from a different subject. Colors indicate the mean diffusion orientation within each cluster. Expert labels are overlaid as abbreviations from Table 4.1 on one of the thalami.

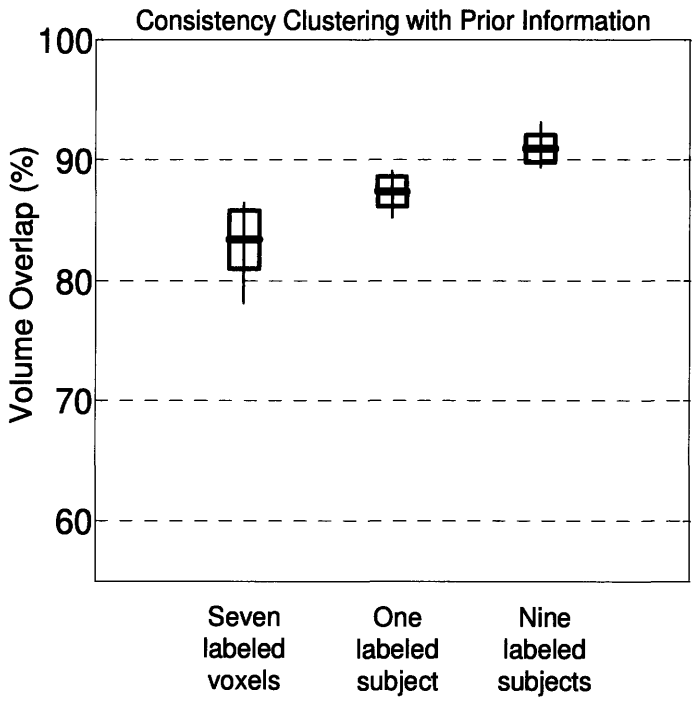

Figure 5-11: Quantitative evaluation of Consistency Clustering with varying amount of prior information. The most accurate segmentation from Figure 5-8 is presented for the "one labeled subject" experiment. The boxes indicate one standard deviation around the mean, and the thin lines indicate the range.

5.4.4 Comparison of Segmentation Methods

To compare Consistency Clustering with the previous segmentation methods presented in Chapter 4, each subject's thalami were segmented individually using the k-means algorithm as described in [81] with uniform initialization and using the spectral clustering algorithm presented in Section 4.4. The segmentation results from these experiments are presented in Figure 5-12 and Figure 5-13, respectively. When compared with Figure 5-10, the lack of consistency among subjects becomes clear for both methods. This lack of consistency is also apparent in the volume overlaps, which are presented for different methods in Figure 5-14.

Not surprisingly, Consistency Clustering with prior information performed better than the unsupervised methods. The worse performing method was the k-means algorithm of [81], indicating the need for prior information and the weakness of the unsupervised algorithms for replicating expert preference.

The Consistency Clustering took under 2 minutes for each case to converge on a desktop personal computer with a non-optimized MATLAB implementation for the joint segmentation of 10 subjects. The algorithm's complexity is linear with the number of voxels for fixed number of clusters.

Figure 5-12: Segmentation results from k-means algorithm. Each image depicts a segmentation result obtained from a different subject. Colors indicate the mean diffusion orientation within each cluster.

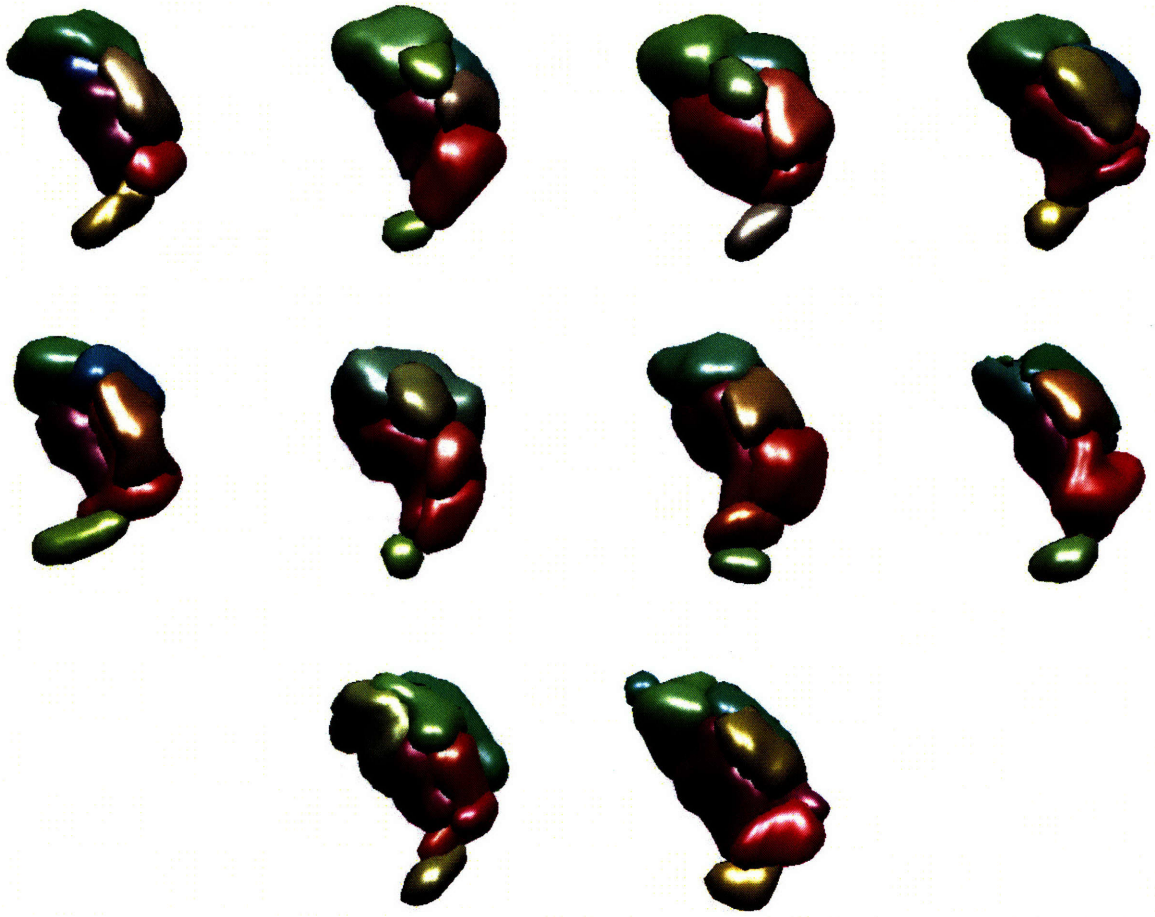

Figure 5-13: Segmentation results from spectral clustering algorithm. Each image depicts a segmentation result obtained from a different subject. Colors indicate the mean diffusion orientation within each cluster.

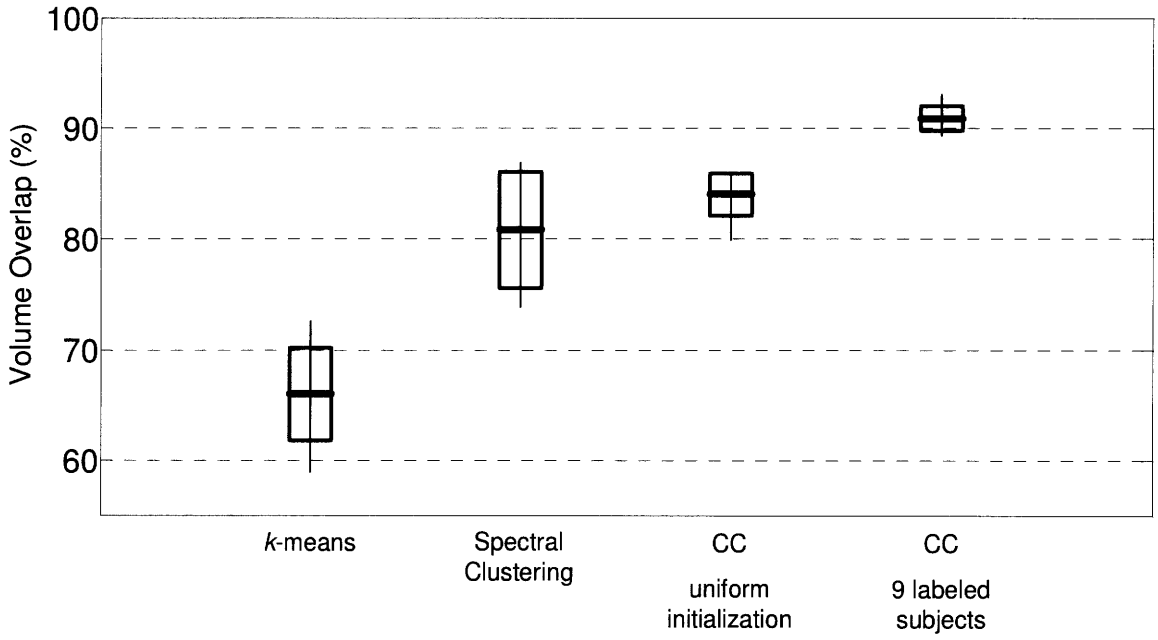

Figure 5-14: Quantitative evaluation of the k-means, spectral clustering and Consistency Clustering. The boxes indicate one standard deviation around the mean, and the thin lines indicate the range.

5.5 Summary of the Chapter

In this Chapter we presented a novel algorithm, called Consistency Clustering, for jointly segmenting a population of diffusion tensor images of the deep gray matter. The joint segmentation resulted in highly accurate segmentations for each subject as well as correspondence between the subjects. This is an important difference between Consistency Clustering and previous algorithms proposed to segment the gray matter, since without correspondence between the segmentations of individual subjects, it is difficult to assign consistent anatomical labels to the resulting segmentations. Also, without consistent and anatomically meaningful segmentations, the quantitative morphometry becomes a challenge in the gray matter.

Consistency Clustering provided consistent segmentations for the population, either with prior information about the expert labels or in a completely data-driven fashion. Also, through the use of labels from other subjects in the population, the algorithm was able to produce segmentations that were both qualitatively and quantitatively very similar to the expert's preference. Therefore, Consistency Clustering can be used in two different ways to produce consistent segmentations in a population. The first way involves running the algorithm unsupervised on the population, and then assigning anatomical labels to the segmentations only on one of the subjects. The labels are then automatically transferred to the rest of the subjects since the correspondence problem is already solved at this stage. The second way involves labeling one or several subjects by hand, and then using these labeled subjects as prior information to label the rest of the population according to the expert preference.

Consistency Clustering learns the parameters of a common probabilistic model for the population from the DTI data. The accuracy of the segmentations depend on the effectiveness of the model in representing the data. In this Chapter, we used a simple parametric data model that included a spatial Gaussian distribution and a von-Mises Fisher distribution for the principal diffusion direction. Furthermore, we made simplifying assumptions for the observed data, such as independence between the spatial observations and the principal diffusion direction for a given voxel and independence between the spatial locations

of voxels. Neither of these assumptions are realistic; however, they result in less complexity for the algorithm. We also used a poly-rigid registration framework, which is also a simple model for registration problem and results in a simpler algorithm. Consistency Clustering with this simple model resulted in over 92% volume overlap between the expert labeled data and the automatic segmentations, indicating that the model is sufficient for accurate segmentation of the DTI data.

In this Chapter we compared our automatic segmentation results against labels produced by a single expert, and reported 92% volume overlap between the expert labeled data and the automatic segmentations when performed using prior information. In general, comparision with a single expert is not desirable, since there is inevitably some variance between labels obtained by different experts. However, additional expert labeled datasets for thalamic nuclei are not available at this time. Even then, we can compare our results with another automatic segmentation method proposed by Pohl [62], where Pohl compared automatic segmentation results of the whole thalamus obtained from structural MRI against an expert labeling. In that case, Pohl reported 89% volume overlap between the automatic segmentations and an expert labeled image. Futhermore, Pohl evaluated the automatic segmentation algorithm results (from a different structure) against the reference standard obtained by the STAPLE algorithm, which aims to generate a ground truth segmentation from multiple expert labeled images [75]. Pohl's segmentation results were comparable to all six experts, and in fact better than two of the experts. Therefore, we believe the 92% volume overlap we report would be comparable to an expert labeling, and it could be tested if multiple expert labeled data were available for the thalamic nuclei.

Either completely data-driven or with prior information, Consistency Clustering (or a variant with an improved model for the thalamic nuclei) is a powerful tool that provides fast and consistent segmentation of the deep gray matter and has a use in a variety of applications such as in quantitative morphometry studies and pre-surgical planning.

Chapter 6

Consistent White Matter Segmentation

In this chapter, we extend Consistency Clustering to automatically construct a probabilistic white-matter atlas from a set of multi-subject diffusion weighted MR images. We formulate the atlas creation as a maximum likelihood problem which the proposed method solves using a generalized Expectation Maximization (gEM) framework. Additionally, the algorithm employs an outlier rejection and denoising strategy to produce sharp probabilistic maps of certain bundles of interest. We test this algorithm on synthetic and real data, and evaluate its stability against initialization. We demonstrate labeling a novel subject using the resulting spatial atlas and evaluate the accuracy of this labeling. Consistency Clustering is a viable tool for completely automatic white-matter atlas construction for subpopulations and the resulting atlas is potentially useful for making diffusion measurements in a common coordinate system to identify pathology related changes or developmental trends.

6.1 Introduction

The human brain consists of a large number of distinct structures along with a complex network of white matter tracts connecting these structures. Neuroscientists have long been interested in studying these white matter connections, and have advanced their understanding through histological studies of sliced animal brains [36] and by studying single tracts through the injection of markers followed by a 3D imaging experiment [45]. However, histological studies are limited to analysis of highly deformed 2D slices and thus are not adequately informative of the complex 3D connections, and marker studies are limited to imaging a single tract at a time. This is why the recent introduction of diffusion weighted imaging (DWI) as a new magnetic resonance imaging (MRI) modality caused much excitement.

DWI is an imaging modality that measures free water diffusion, i.e. Brownian motion, of the endogenous water in tissue. This water diffusion is measured for the entire brain using directional gradients in the MRI machine. Unlike structural MRI, however, DWI does not result in a single image, but a series of images, each one quantifying the amount of water diffusion in a specific direction. These images are then combined together to calculate a 3D water diffusion probability distribution function (pdf) for every voxel. The most common way of calculating this pdf is to assume the water diffusion is Gaussian for a given voxel and fit the data from individual DWI images to this model, resulting in the diffusion tensor images (DTI) [6].

In human brain tissue, the water diffusion is not the same in all directions, since it is obstructed by structural elements such as cell membranes or myelin. When this obstruction constrains the water diffusion in a coherent direction, such as within the cerebral white matter, the resulting water diffusion tensor becomes anisotropic, containing information about the directionality of the axon bundles [43, 61]. Thus, quantification of water diffusion in tissue through DTI provides a unique way to look into white matter organization of the brain.

With the observation that the diffusion within the white matter is highly directional, an interesting area of research has emerged, aimed at reconstructing the underlying white

matter architecture from DTI data. Reconstruction typically consists of starting at a voxel and following the most likely path of diffusion in small steps, until a stopping criteria is reached. This sort of reconstruction is called tractography and results in a 3D curve, called a fiber [7]. Note that these fibers should not be confused with individual axons since we know that typical in-vivo imaging resolution is too low to reconstruct individual axons from the DTI images. However since these axons tend to be situated coherently in space, the larger white matter tracts appear as fiber bundles in tractography [46, 47, 56]. If tractography is performed by seeding in every white matter voxel in the brain, the resulting set of fibers spans the whole brain.

A natural extension of tractography is to group fibers into bundles and assign them anatomical labels. One popular technique to construct fiber bundles is to manually trace a region of interest (ROI) that is thought to correspond to a certain anatomical structure, and initialize the tractography algorithm from the set of voxels within this ROI [13]. The resulting tracts are then grouped together into a fiber bundle. A slightly more advanced variant of this technique involves seeding at every voxel in the brain that is thought to belong to the white matter, and grouping all the fibers that pass through the previously traced ROI into a bundle [83]. It is also possible to use multiple ROIs and it has been reported that this approach improves consistency between subjects [28].

ROI based methods are popular for their flexibility but could be subject to user bias since ROIs are manually traced. Therefore several methods have been proposed to identify anatomically meaningful regions from the DTI data without dependence on manually traced ROIs. One class of methods automatically groups tractography results, which were seeded everywhere in the brain, into bundles through clustering algorithms, e.g. [46, 47, 56].

Even though resulting fiber bundles from the aforementioned algorithms, in general, match well with the known anatomy, the quality of the results suffer from low SNR and low resolution of diffusion images. These issues lead to at least two types of artifacts in tractography: early termination of fibers and deviation of fibers from one bundle to another [83] (See Figure 6.1 for the latter type of artifacts).

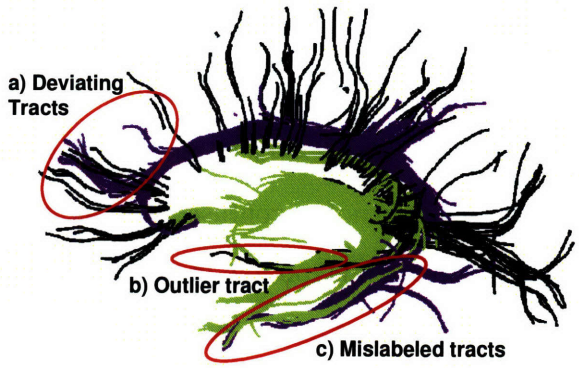

Figure 6-1: Some common tractography and clustering errors. Tracts from the fornix (in green) and the cingulum (in purple) bundles along with a few selected tracts from the corpus callosum (in black) are shown as labeled using the high dimensional atlas of Section 6.2.4. The tractography noise is evident in the image as tracts deviating from one bundle to another (a) as well as a few outlier tracts (b). Also, these images contain instances where the high dimensional atlas failed to label the tracts correctly (c).

In an attempt to reduce such errors in tractography, there have been some studies that repeat the same analysis on several individual subjects and average the results to create group maps of white matter tracts [83, 14]. By cross-subject averaging, the goal is to create a group map that is more robust to multiple sources of inconsistency: noise in tractography, inter-subject anatomical variability and registration errors. These methods involve single subject ROI drawing, followed by bundling of tracts that pass through these ROI's. This yields a binary image for each bundle in each subject, that represents whether a fiber from the bundle of interest is present or not in that voxel. The next step is to employ an "off-the-shelf" registration algorithm that ignores DTI information to co-register subjects into a common coordinate frame. This is followed by an averaging of the binary images.

In this Chapter we combine the group-wise registration and tract labeling with the atlas construction into an integrated and principled method. This method segments corresponding white-matter regions from a group of subjects while creating a spatial white-matter atlas of the common structures for a sub-population in a completely automatic and data-driven fashion. For this purpose, we start with a group-wise clustering to identify fiber bundles in individual subjects as well as correspondences among subjects. That is followed by a novel learning algorithm, which we call "Consistency Clustering". This algorithm utilizes generalized Expectation Maximization (gEM) [51] for optimization. Consistency Clustering

performs group-wise labeling, de-noising, outlier rejection, and group-wise registration in an iterative, integrated scheme and results in a spatial map that represents the most common structures in a group of subjects.

The possible uses for this algorithm are multi-fold. It could be used to compute a spatial atlas that retains only the most reproducible tracts. This atlas would represent consistently present white matter tract regions in a group of subjects. Knowing what is normal also provides a way to identify what is not, which is useful for pathology identification. Also of interest is a statistical framework for the comparison of different sub-populations, to identify disease-related or developmental changes.

6.2 Data Collection and Pre-Processing

In this section, image acquisition details and the tools used for preprocessing the DTI data are introduced. These tools include diffusion tractography and the creation of a high dimensional atlas through spectral clustering of the tractography results.

6.2.1 Image acquisition

MRI data were acquired on fifteen healthy young participants at Johns Hopkins University. The data were acquired on a 1.5 Tesla Philips Medical Systems MRI scanner using an 8-channel head coil. All participants provided informed written consent by the guidelines of the Johns Hopkins University Internal Review Board. The participants gave their informed consent in writing prior to the session.

The DTI data were acquired using a single-shot spin-echo EPI sequence. The sequence parameters were TR/TE=7000/80 ms, b=700 s/mm^2, 5 T2 images, 30 diffusion gradient directions, 1 average, with total acquisition time 12-15 minutes. Fifty to sixty axial slices were acquired. The field-of-view was 256×256 mm and the matrix size was 96×96 to give 2.5×2.5 mm in-plane resolution. The slice thickness was 2.5 mm with 0 mm gap.

6.2.2 Tractography

Diffusion tensor imaging models microscopic diffusion in the brain tissue as a Gaussian diffusion process and encodes this information in a 3×3 diffusion tensor that is analogous to the covariance matrix of a 3-dimensional Gaussian random variable. The principal eigenvector of this tensor is aligned with the most likely direction of diffusion and that is the direction a streamline tractography algorithm aims to follow. In this work, tractography was performed using a Runge-Kutta order two integration of the underlying tensor field. The algorithm is initiated in every voxel in the brain with a C_L value exceeding 0.25, since having a high C_L value is an indicator of coherent diffusion [79]. The quantity C_L is defined as $\frac{\lambda_1-\lambda_2}{\sqrt{\lambda_1^2+\lambda_2^2+\lambda_3^2}}$, where $\lambda_1, \lambda_2, \lambda_3$ are the eigenvalues of the diffusion tensor sorted in descending order [79]. Tractography is performed at a step size of $0.5\ mm$ until the trajectory comes to the edge of the white matter, defined by a threshold of $C_L = 0.15$.

6.2.3 Initial Registration

The group-wise clustering tool we will employ assumes all subjects' tractography is in a common coordinate frame, therefore a spatial normalization is needed. This normalization is performed using a group-wise, template-free affine registration algorithm [87], known as "congealing", on the Fractional Anisotropy (FA) images generated for each subject. This initial normalization aims to remove gross differences across subjects due to global head size and orientation. It is thus limited to a 9 parameter affine transformation that accounts for scaling, rotation and translation. The resulting transformations are then applied to each of the computed fibers to map them into a common coordinate frame for clustering.

6.2.4 Fiber Clustering for a High Dimensional Atlas

Fiber clustering is based on the assumption that fibers that are in the same bundle follow a similar trajectory within the tissue. To quantify this similarity, the employed algorithm uses the mean closest point distance [17]. This is defined as the average distance from each point in one tract to the nearest point on the other tract. This distance is then symmetrized by taking the minimum of the two distances between every fiber pair. The algorithm converts this distance into an affinity measure using a Gaussian kernel. Once these affinities are calculated for every tract pair, we employed spectral clustering with normalized cuts criteria [67] to compute clusters of fibers, similar to the voxel-based segmentation of Chapter 4.

Once clustering is completed, an expert needs to assign anatomical labels to these clusters. This is done interactively on one of the subjects by selecting clusters one by one and assigning them anatomical labels. Since the clustering is done on a number of subjects jointly, these labels are automatically transferred to all the subjects used for the clustering stage.

By clustering fibers from multiple subjects into bundles, common white matter structures are discovered in an automatic way, and the bundle models are saved with expert anatomical labels to form a high dimensional atlas [58]. This atlas is called high dimensional since the clustering is done in a high dimensional space and the cluster definitions as well as the anatomical labels are saved in this high dimensional space. Even though re-

sulting fiber bundles from labeling with such an atlas in general match well with the known anatomy, the quality of the results exhibit artifacts such as mislabeled tracts and outlier tracts labeled as with one of the anatomical labels (See Figure 6.1). Also the quality of the spatial atlas obtained by simply voxelizing a set of multi-subject images labeled with a high dimensional atlas is limited (see speckles in Figure 6-6a), which indicates that automatic identification of consistent bundles and their localization in the common coordinate frame is problematic due to limited inter-subject alignment, cross-subject anatomical variability and tractography noise.

In the current work, we utilize the automatically segmented tractography results obtained through the aforementioned semi-supervised, high-dimensional atlas framework as initial but noisy tract labels, and perform a group-wise labeling, de-noising, outlier rejection, and registration in an iterative and integrated scheme while building a spatial atlas that represent the most common structures.

6.3 Consistency Clustering for White Matter Segmentation

In this Section, we introduce several tools to discover the most consistent bundles across subjects and obtain a sharp atlas of these bundles. These tools include a nonlinear registration component defined on the bundles, outlier rejection and a "tract cut" operation. We formulate the problem as maximum likelihood and solve it using a generalized EM algorithm. The probabilistic framework is well-suited for extensions such as inclusion of anatomical prior information and integration with existing probabilistic atlases based on expert labeled structural MRI.

Probabilistic atlas building as described in [83] involves a sequential process of

1. Registrating a set of DTI images,

2. Labeling tracts through manually drawn ROI's

3. Stacking the registered labels and statistical averaging; i.e., counting the number of times each voxel inherits the same label.

One of the premises of such an atlas is its ability to label registered tracts. Consistency Clustering we propose in this work capitalizes on this premise by iteratively relabeling tracts while building the atlas. This iterative process has the advantage of ensuring that tract labels match with the atlas that is being built and therefore results in a sharper and more consistent result. (See Figure 6-6c).

We start with a set of subjects indexed with $s \in \{1, 2, \ldots, S\}$ and a set of tracts from each subject, $T = \{T_s\}$ and $T_s = \{t_i\}$. We formulate our problem as a maximum likelihood

estimation of the parameter set Θ:

$$\begin{aligned}
\Theta^* &= \arg\max_{\Theta} \prod_{s=1}^{S} p(T_s|\Theta) \\
&= \arg\max_{\Theta} \sum_{s=1}^{S} \log p(T_s|\Theta) \\
&= \arg\max_{\Theta} \sum_{s=1}^{S} \sum_{t \in T_s} \log p(t|\Theta) \\
&= \arg\max_{\Theta} L(\Theta|T)
\end{aligned}$$

The term Θ includes the parameters we wish to estimate and will be defined later to include a non-parametric spatial mixture density and registration parameters. Here we assumed the tracts are statistically independent from each other in calculating $\mathscr{L}(\Theta)$, the log likelihood of the data, i.e. the observed tracts. We can then derive the Expectation Maximization update equations by marginalizing this likelihood term over a set of cluster (in this context, bundle) labels $c \in \{1, 2, \ldots, C\}$,

$$\begin{aligned}
\mathscr{L}(\Theta) &= \sum_{s=1}^{S} \sum_{t \in T_s} \log \sum_{c=1}^{C} p(t, c|\Theta) \\
&= \sum_{s=1}^{S} \sum_{t \in T_s} \log \sum_{c=1}^{C} p(t, c|\Theta) \frac{q(c)}{q(c)} \\
&\quad \text{s.t.} \sum_{c=1}^{C} q(c) = 1,
\end{aligned}$$

for any distribution $q(c)$, defined on the cluster labels. Now, using Jensen's inequality:

$$\mathscr{L}(\Theta) \geq \sum_{s=1}^{S} \sum_{t \in T_s} \sum_{c=1}^{C} q(c) \log p(t, c|\Theta) + \text{const} \tag{6.1}$$

$$= \sum_{s=1}^{S} \sum_{t \in T_s} \mathbb{E}_q \left[\log p(t, c|\Theta)\right] + \text{const.} \tag{6.2}$$

The constant term does not depend on Θ and $\mathbb{E}_q$ denotes expectation with respect to q. Further, the equality of Equation (6.1) is met if and only if $q(c) = p(c|t, \Theta^{(n)})$, for a given

(fixed) $\Theta^{(n)}$ [51]. The generalized EM algorithm maximizes this lower bound using the current estimate of the parameters, $\Theta^{(n)}$, in the E-step and re-estimates the parameters in the M-step. This iterative optimization is guaranteed to converge to a local optimum of the parameter set, Θ^*. Now, let's define the probability of a tract through a non-parametric and voxelized spatial mixture density distribution:

$$\begin{aligned} p(t|\Theta) &\triangleq \sum_{c=1}^{C} \pi_c p(t|c,\Theta) \\ &= \sum_{c=1}^{C} \pi_c \prod_{x \in R_s(t)} \theta_c(\lfloor x \rfloor), \end{aligned}$$

where π_c are the mixture weights and $\sum_c \pi_c = 1$. $R_s : \mathbb{R}^3 \mapsto \mathbb{R}^3$ is a subject-specific spatial transformation and $R_s(t)$ denotes a tract obtained by applying the spatial transformation to a tract t. $x \in \mathbb{R}^3$ are the samples along a given tract and $\lfloor x \rfloor$ denotes quantization of x, i.e., the voxel that sample falls into. And, finally, $\theta_c(\lfloor x \rfloor)$ represents a non-parametric spatial distribution, quantifying the spatial probability of observing tracts from cluster c. Naturally, $\sum_{x \in X} \theta_c(x) = 1$ for all $c \in C$. Then, the parameter set we wish to estimate becomes, $\Theta = \{\theta_c(\lfloor x \rfloor)\} \cup \{\pi_c\} \cup \{R_s\}$.

In the next sections we derive the specific update equations for our formulation to iteratively estimate Θ^*.

6.3.1 E-Step

In the E-step, the algorithm updates the membership probabilities for each of the tracts given the current estimate of the parameter set $\Theta^{(n)}$.

$$\begin{aligned} p(c|t,\Theta^{(n)}) &\propto p(t|c,\Theta^{(n)}) p(c|\Theta^{(n)}) && (6.3) \\ &\propto \pi_c^{(n)} \prod_{x \in t} \theta_c^{(n)} \left\lfloor R_s^{(n)}(x) \right\rfloor && (6.4) \\ &\triangleq \tilde{p}_{tc}^{(n)}, && (6.5) \end{aligned}$$

where, at every iteration (n), $\sum_c \tilde{p}_{tc}^{(n)} = 1$ for all tracts t.

6.3.2 M-Step

In the M-step, the algorithm updates the parameter set Θ to maximize the expected value of the log likelihood. From Equation (6.2), ignoring the constant term which is independent of Θ, Θ^* is computed as:

$$\begin{aligned}
\Theta^* &= \arg\max_{\Theta} \sum_{s=1}^{S} \sum_{t \in T_s} \sum_{c=1}^{C} \tilde{p}_{tc} \log p(t, c|\Theta) \\
&= \arg\max_{\Theta} \sum_{s=1}^{S} \sum_{t \in T_s} \sum_{c=1}^{C} \tilde{p}_{tc} \log \left[\pi_c \prod_{x \in t} \theta_c(\lfloor R_s(x) \rfloor) \right] \\
&= \arg\max_{\Theta} \sum_{s=1}^{S} \sum_{t \in T_s} \sum_{c=1}^{C} \tilde{p}_{tc} \left[\log \pi_c + \sum_{x \in R_s(t)} \log \theta_c(\lfloor x \rfloor) \right] \qquad (6.6) \\
&\quad \text{s.t. } \sum_x \theta_c(x) = 1 \text{ , for all } c \text{ and } \sum_c \pi_c = 1
\end{aligned}$$

For a fixed parameter set $\Theta^{(n)}$ the update equations for $\pi_c^{(n+1)}$ and $\theta_c^{(n+1)}(x)$ can be derived using Lagrange multipliers for the corresponding constraints and setting the derivative of Equation (6.6) to zero. The resulting update equations are:

$$\pi_c^{(n+1)} \propto \sum_{s=1}^{S} \sum_{t \in T_s} \tilde{p}_{tc}^{(n)} \qquad (6.7)$$

$$\theta_c^{(n+1)}(x) \propto \sum_{s=1}^{S} \sum_{t \in T_s} \tilde{p}_{tc}^{(n)} N(\lfloor x \rfloor \in \lfloor R_s^{(n)}(t) \rfloor) \qquad (6.8)$$

where $N(\lfloor x \rfloor \in \left\lfloor R_s^{(n)}(t) \right\rfloor)$ denotes the number of times warped tract samples $\left\lfloor R_s^{(n)}(t) \right\rfloor$ appear in voxel $\lfloor x \rfloor$. The normalizing constants are computed so that $\sum_c \pi_c^{(n+1)} = 1$, and $\sum_x \theta_c^{(n+1)}(x) = 1, \forall c$.

Registration

Unlike $\pi_c^{(n+1)}$ and $\theta_c^{(n+1)}(x)$, the registration parameters $R_s^{(n+1)}$ do not have an analytical solution. However, we derive the maximum likelihood optimization function, and optimize that iteratively using a numerical scheme. Then for each subject, the registration parameters

are updated as $R_s^{(n+1)}$:

$$
\begin{aligned}
R_s^{(n+1)} &= \arg\max_{R_s} \sum_{t \in T_s} \sum_{c=1}^{C} \tilde{p}_{tc}^{(n)} \sum_{x \in R_s^{(n)}(t)} \log \theta_c^{(n+1)}(\lfloor x \rfloor) \\
&= \arg\max_{R_s} \sum_{t \in T_s} \sum_{c=1}^{C} \sum_{x \in R_{sc}^{(n)}(t)} \tilde{p}_{tc}^{(n)} \log \theta_c^{(n+1)}(\lfloor x \rfloor) \qquad (6.9)
\end{aligned}
$$

In the last equation we parameterize the registration R_s as a set of affine transformations $\{R_{sc}\}$ for each subject s and for each cluster c, and solve for these separately. We perform this optimization using a simplex search method [40] to find the 9 affine parameters that correspond to translation, rotation and scaling (no shearing).

6.3.3 Tract Cut

Even with the correct labeling and perfect registration, we will not be able to recover from the errors in tractography that are present as deviations of tracts from one bundle to another. One way to reduce the effects of this sort of artifacts is to remove the tract samples that do not agree with their corresponding tract labels. Our problem formulation offers a natural way to identify these deviating tract samples since it offers a maximum likelihood estimate for the label of each tract sample as well as the whole tract. The segments of every tract that consist of samples whose maximum likelihood label do not agree with the tract's overall label are identified as deviating tract segments and can be separated from the rest of the tract, resulting in a shorter tract whose every sample has the same maximum likelihood label (See Figure 6-2). More specifically, the tract cut operation removes tract samples from the tips of each tract, if the samples' labels do not match with the tract's overall label. The removal starts at the two ends of each tract, and continues until the label of the remaining sample at the tip of the shortened tract is no longer inconsistent with the rest of the tract.

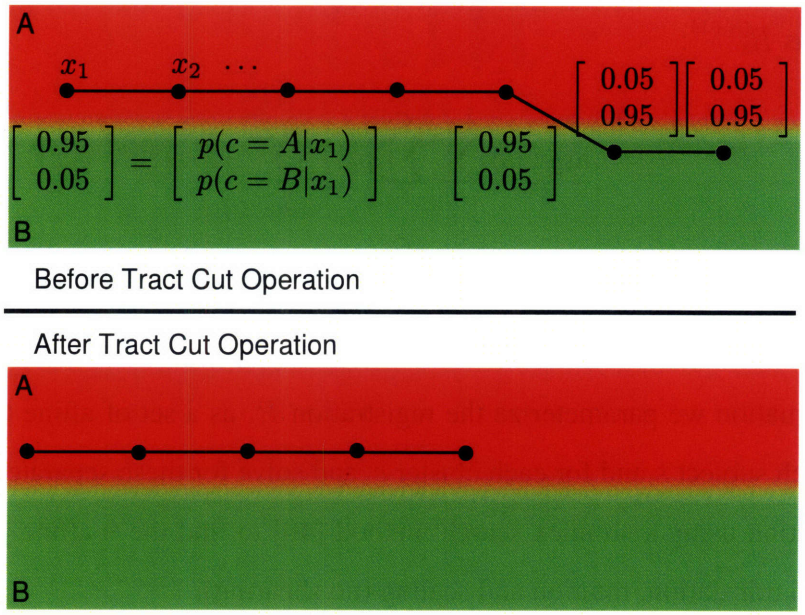

Figure 6-2: Illustration of the tract cut operation. The atlas contains two labels: A (in red) and B (in green). Each sample of the tract has been assigned a probability vector, $p(c|x_i)$ for $c \in \{A, B\}$. The overall tract label is determined as the one that maximizes the tract likelihood: $p(c|t) \propto p(c) \times \prod_i p(x_i|c)$. However, there are parts of this tract that does not agree with the tract's overall label. Tract cut operation removes these parts of the tracts.

6.3.4 Outlier Rejection

Another type of artifact is the presence of tracts that do not belong to any of the clusters. These are considered to be outlier tracts. Since these tracts are situated in the areas that are unlikely to belong to any of the clusters that are under investigation, they can be automatically identified and removed with the introduction of a default label that has a fixed very low probability throughout the image (See Figure 6-3).

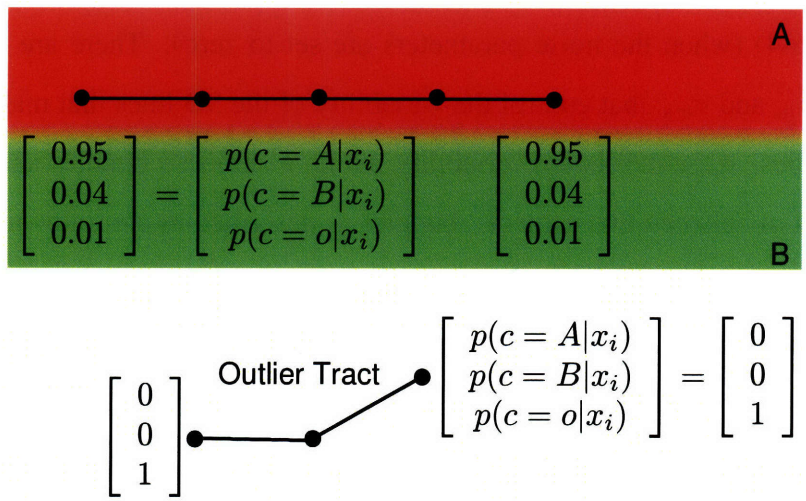

Figure 6-3: Illustration of outlier removal. The atlas contains two labels: A (in red) and B (in green). A third label is created for the outliers. Each sample of the tract has been assigned a probability vector, $p(c|x_i)$ for $c \in \{A, B, o\}$, where 'o' is the label for outliers. The overall tract label is determined as the one that maximizes the tract likelihood: $p(c|t) \propto p(c) \times \prod_i p(x_i|c)$. If this likelihood is greater for the outlier class than any other, the tract is labeled as an outlier.

6.4 Experiments

In this section we demonstrate the dual purposes of the Consistency Clustering algorithm: a group-wise clustering algorithm, as well as a robust atlas building algorithm. We apply the algorithm on synthetic data and in-vivo DTI data.

Synthetic data is created to simulate 5 subjects with two fiber bundles in each subject. The fiber bundles cross each other in the middle of the data range at a 30 degree angle (See Figure 6-4) (when the noise parameters are set to zero). There are also two noise parameters, σ_{in} and σ_{btw} that control the variability of the 3D lines that make up each one of these bundles. σ_{in} controls the variability within a subject's bundles and σ_{btw} controls the variability in between two subjects' bundles as described in the following: For a given starting point x_{st} and a norm-one slope vector s_i for $i = 1, 2$ - one for each bundle, each subject k is assigned a random starting point $x_{st}^k = x_{st} + N(0, \sigma_{btw})$, and a random slope vector $s_i^k = s_i + N(0, 0.2 \times \sigma_{btw})$, where $N(0, \sigma)$ is a zero-mean 3 dimensional random variable with a diagonal covariance matrix whose elements are set to σ. Once each subject's starting point and slope is determined, each tract within the subject is generated by randomly perturbing these values to $x_{st}^k + N(0, \sigma_{in})$ and $s^k + N(0, 0.2 \times \sigma_{in})$. Each line then consist of a set of uniformly spaced points in 3D, generated using these parameters. The lines are constrained to be within a pre-determined range to limit the length.

Diffusion MRI images are acquired on a 1.5 T Philips scanner with SENSE parallel imaging along 30 noncollinear gradient directions, using a single-shot spin echo EPI sequence along with five non-diffusion-weighted T2 images at a resolution of 2.5 x 2.5mm x 2.5mm per voxel. The b-value was set to $b = 700s/mm^2$. 50-60 slices per subject covering the entire hemispheres and the cerebellum were analyzed for 15 subjects. Tractography was performed in each subject using Runge-Kutta order two integration, with the following parameters: seeding threshold of $C_L = 0.25$, stopping threshold of $C_L = 0.15$, step size of 0.5mm, and minimum total length of 25mm. Initial labels to the resulting tracts are assigned using a high dimensional atlas as described in Section 6.2.4.

6.4.1 Consistency Clustering: A Clustering Algorithm

Consistency Clustering is an atlas-based group-wise clustering algorithm. To demonstrate its use as a clustering algorithm we utilized synthetic data, for which we have the ground truth. To understand the effect of noise on Consistency Clustering, we created synthetic data with varying noise levels. Each synthetic data set consisted of 5 subjects with two fiber bundles for each subject. We initialized Consistency Clustering with random labels as in Figure 6-4 (A and B) and repeated the experiment at each noise level 30 times. The algorithm recovered the ground truth labels with high accuracy for a wide range of noise parameters as presented in Figure 6-4 (C and F).

Unlike synthetic data, real DTI data does not offer ground truth to evaluate our algorithm against. However, we can still evaluate the stability of our algorithm for initialization. For this purpose, we applied Consistency Clustering to 15 subjects without the outlier rejection and the tract cuts (since removing tracts and tract segments complicate the error measurement). Once the algorithm converges on an optimal set of tract labels, we stored these optimal labels to test the stability of our algorithm. A stable algorithm is expected to converge back to the same optimal labels from a range of differing initializations. To evaluate this kind of stability, we randomly changed a certain percentage of the optimal labels (varying between 5% and 50%), and reran the algorithm initialized with these sub-optimal labels. We repeated this experiment 10 times for each percentage level, which is varied in increments of 5%. The results from these experiments are presented in Figure 6-5. For a wide range corruption levels, the algorithm produced nearly identical labels as the initial optimal solution.

6.4.2 Consistency Clustering: An Atlas-building Algorithm

As mentioned earlier, Consistency Clustering is an atlas-based group-wise clustering algorithm and can also be used for creation of probabilistic MRI atlases based on DTI tractography clustering. The corpus callosum, the cingulum and the fornix were selected for investigation in this study, because of the specific challenges they present. These three structures are in close proximity with each other, and that results in many mislabeled fibers

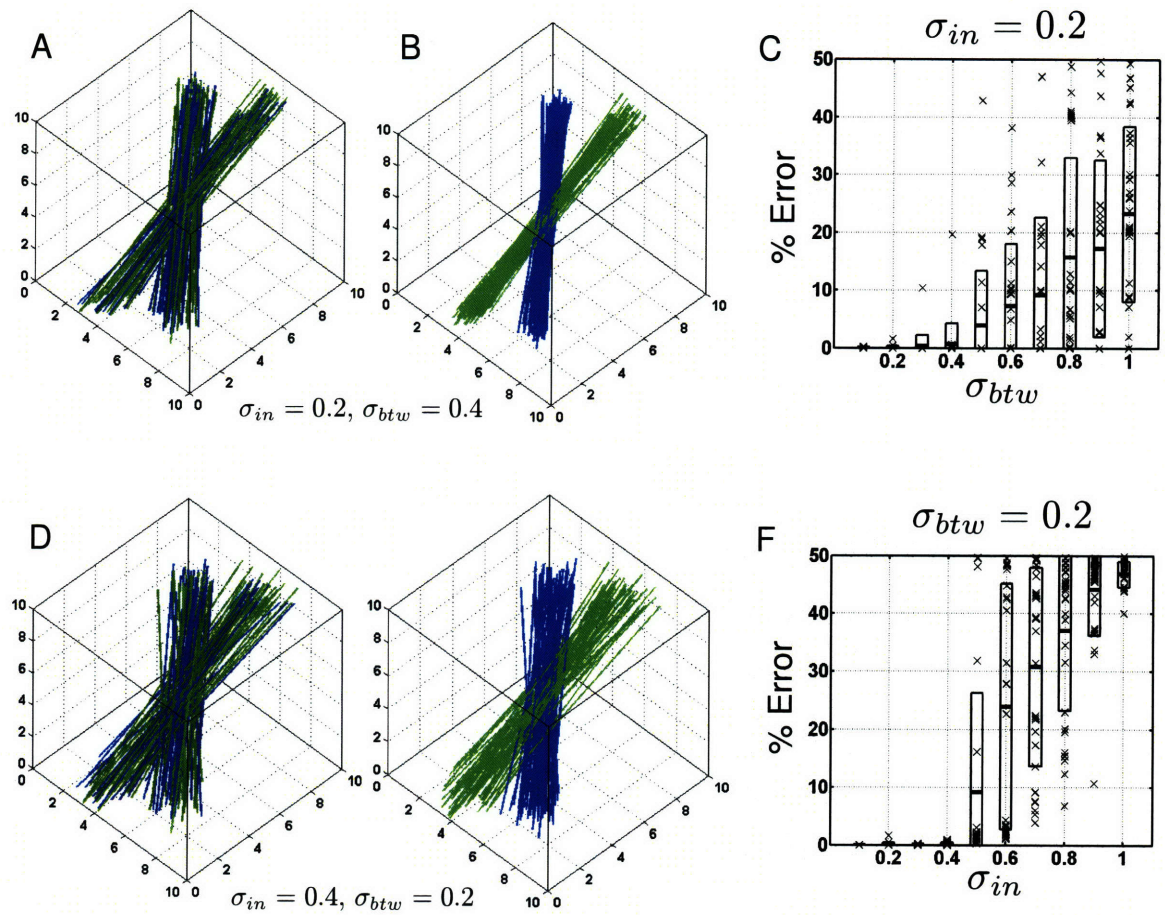

Figure 6-4: Consistency clustering results from a synthetic dataset. Each dataset consists of 5 simulated subjects with two fiber bundles crossing each other at a 30 degree angle. σ_{in} controls the variability of lines within each subject and σ_{btw} controls the variability of lines between the subjects (See Section 6.4 for details). Figures A and D show simulated fibers from 5 subjects overlaid on top of each other and colored according to the initial random labels. Figures B and E show the same datasets as in figures A and D respectively, with the labels and registration parameters resulting from Consistency Clustering. Figures C and F show the algorithm's accuracy in labeling fibers under varying noise levels. The data is randomly created and clustered with random initialization for 30 times at each noise level increment. The cross marks represent observed error levels from each one of these experiments, and the boxes indicate one standard deviation from the mean error at each noise level increment. Figure C shows at least one correctly labeled instance at each noise level between the subjects whereas the algorithm starts to fail regularly with increasing noise within the same subject as shown in Figure F.

when labeled using a high dimensional atlas (see Figure 6-7 (left)). Their close proximity also results in a number of trajectories deviating from one structure to another. These are precisely the sorts of artifacts we wish to reduce through learning common spatial distribu-

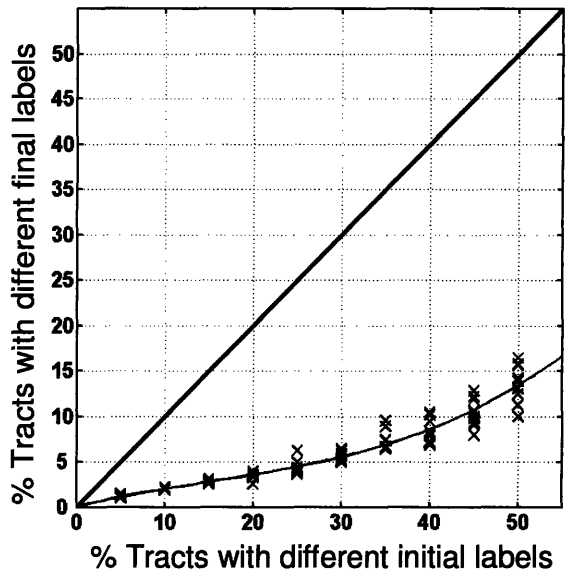

Figure 6-5: Stability of labels produced by Consistency Clustering for the 15 subject DTI data. A set of optimal labels is acquired by running the algorithm on 15 subjects' DTI data until convergence. To test for stability of these optimal labels, a certain percentage of the optimal labels (varying between 5% and 50%) are changed randomly, resulting in sub-optimal labels. These sub-optimal labels are then used as an initialization for the algorithm, to test whether the algorithm would converge back to the same optimal labels. The graph depicts the percentage errors of the tract labels at initialization (x axis) and convergence (y axis) stages of the algorithm. Errors are calculated as the percentage of tracts with non-optimal labels at the corresponding stage. Each cross mark indicates the error from one experiment in which a certain percentage of the optimal labels are changed and the algorithm is rerun with these corrupted labels as initialization. Solid line is produced by fitting a second degree polynomial to these experimental results.

tions of fiber bundles from a group of subjects.

We constructed three different atlases to compare the effects of labeling algorithms on the quality of resulting group maps. The first one is constructed using the initial labels from the high dimensional atlas. A second one is built using Consistency Clustering without the outlier rejection and the tract cut operations, and the last one is generated through Consistency Clustering with the outlier rejection and tract cut operations.

Tracts labeled with the high dimensional atlas qualitatively conformed with the known anatomy of the white matter; however most subjects contained individual tracts deviating from one bundle to another (Figure 6-7 (left)). Also, the high dimensional atlas failed to label some of the tracts correctly, potentially due to registration errors. The proposed

algorithm was able to remove the segments of the tract bundles that were not consistent from subject to subject (Figure 6-7 (right)).

6.4.3 Atlas Quality

Once the atlases are built, it is of interest to quantify the quality of these atlases. We do so through entropy, which is an information theoretic measure of uncertainty in a probability distribution. A smaller entropy measure indicates a sharper atlas, which is due to increasing similarity in the shapes of the bundles that are used to construct that atlas. These relative similarities result from a better registration or more consistent labeling across subjects. The entropy of a spatial distribution of a specific fiber bundle in our atlas notation is defined as:

$$H_c = -\sum_{\lfloor x \rfloor \in x} \theta_c(\lfloor x \rfloor) \log \theta_c(\lfloor x \rfloor)$$

Method	FX	CC	CI
High Dim. Atlas	8.0791	9.6492	8.1250
EM w/o cuts	7.1830	9.5204	7.6818
EM with cuts	6.8434	9.1506	7.0911

Table 6.1: Spatial probability distribution entropies (measured in nats) for each of the structures constructed through three different methods. FX, CC and CI stands for the fornix, the corpus callosum, and the cingulum bundles, respectively. A lower entropy value indicates a sharper distribution that is less contaminated with artifacts. Smaller structures seem to benefit more from the better registration and label correction Consistency Clustering provides.

Removing inconsistencies between subjects before averaging resulted in a qualitatively sharper atlas, which is presented in Figure 6-6. The colored overlays indicate the probability of a tract bundle being present in any location in the brain. A whiter color indicates regions where there were more consistent alignment in the labeled tracts among the subjects, resulting in higher probability values. Spatial probabilities were greatest in the central regions of the bundles under investigation in all three atlases. However, these probabilities increase in the central regions and decrease in the perimeters with improved registration and corrected labels through the proposed algorithms. This resulting improvement in the

atlas quality is quantified in Table 6.1 through the distribution entropies. This analysis indicates that smaller structures benefit more from the better registration and label correction Consistency Clustering provides.

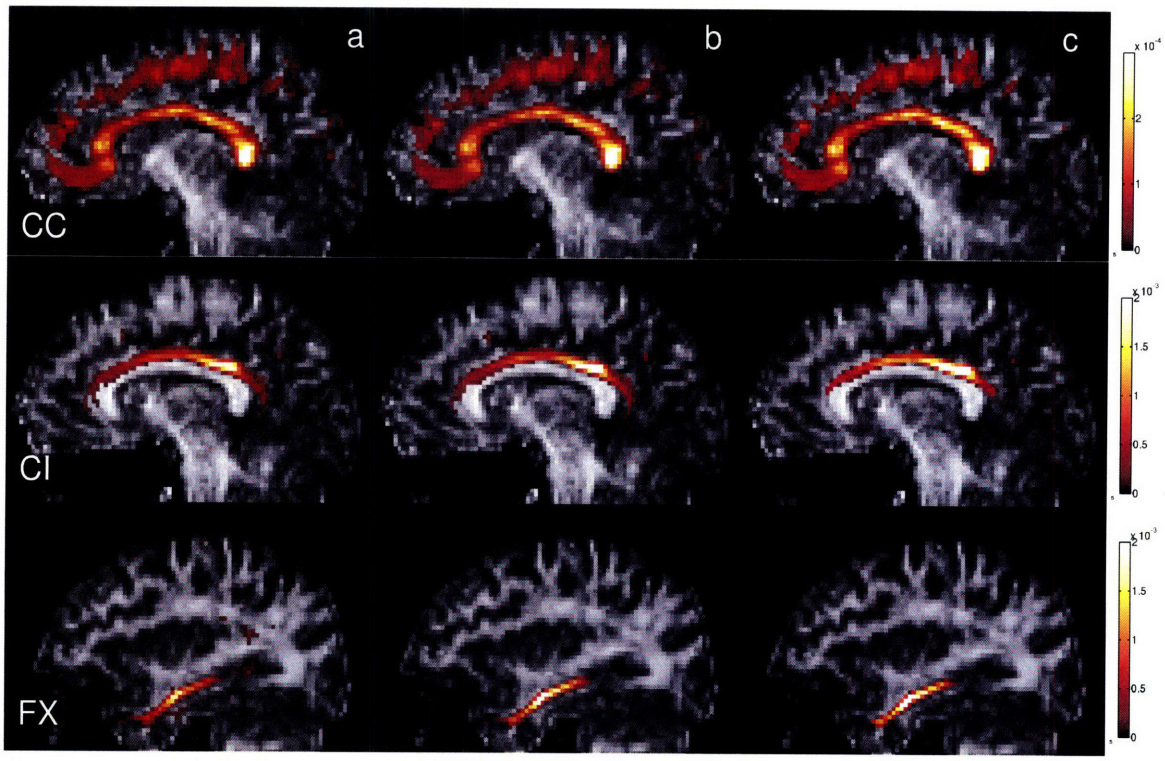

Figure 6-6: Spatial distributions of the corpus callosum, the cingulum and the fornix bundles from three single slices overlaid on their corresponding FA images. These maps are constructed using three different methods. a) High dimensional atlas, b) Consistency Clustering without tract cuts, c) Consistency Clustering with tract cuts. The colorbars indicate the probability of each voxel in the spatial distribution of the corresponding fiber bundle. Note that the probabilities become higher in the central regions of the bundles and the number of sporadical voxels with non-zero probabilities decrease from left to right, indicating a sharper atlas through better registration and more consistent labeling of the subjects. This observation is quantified through these distributions' entropies in Table 6.1.

Another way of quantifying the atlas quality is through its accuracy in labeling novel subjects that were not used to build the atlas. Once again since we do not have the ground truth for the DTI tracts, we ran Consistency Clustering on all 15 subjects without the outlier rejection and the tract cuts and stored the resulting tract labels for later evaluation. Then we repeated this atlas building experiment 15 times, each time with 14 subjects participating in the atlas building stage. Once such an atlas is constructed, the remaining subject is

Figure 6-7: Tracts from the fornix (in green) and the cingulum (in purple) bundles along with a few selected tracts from the corpus callosum (in black) as labeled through the high dimensional atlas (left) and through Consistency Clustering with tract cuts (right). The tractography noise is evident in the images on the left as tracts deviating from one bundle to another. Also, these images contain instances where the high dimensional atlas failed to label the tracts correctly. Consistency Clustering is able identify consistent tract bundles across subjects, while tract cuts operation remove the inconsistent parts of these bundles, resulting in "core" bundles.

labeled using the atlas through a modified EM loop that involves the registration step and the E-step until convergence. Once each one of the subjects are labeled in this way, the resulting labels are then compared against the labels produced when all 15 subjects were used for the atlas construction. The overall differences in the labels were in the range of 1.0% to 2.8%. That translates to more than 97% of the labels being the same whether a subject was used in the atlas construction or not. The results from these experiments are

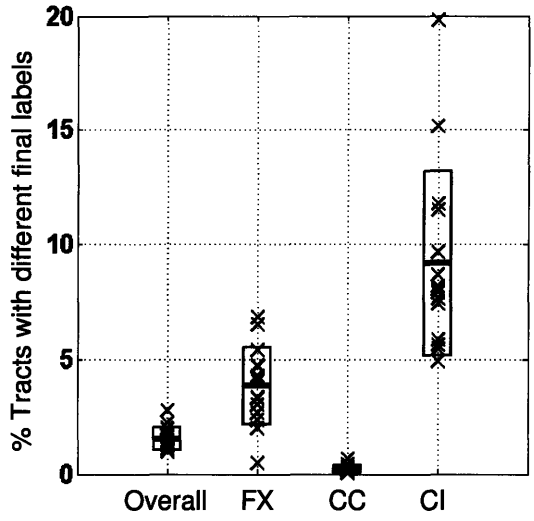

Figure 6-8: Evaluating the consistency of atlas-based labeling. Each cross mark represents an experiment in which one subject labeled with the atlas constructed from the remaining 14 subjects. These labels are compared against the labels produced by the atlas building algorithm with all 15 subjects included. The boxes indicate one standard deviation above and below the mean difference level. Leftmost column shows the overall differences, summarizing the results for the whole atlas, and the the other three columns shows the differences for each of the structures represented in the atlas. These structures are the fornix (FX), the corpus callosum (CC) and the cingulum (CI).

presented in Figure 6-8 for each one of three structures used in the atlas building as well as the overall results. Labels in the cingulum bundle showed more variability than the labels in other structures, due to the high number of tracts that deviate from the cingulum to the corpus callosum. Even though the absolute effect of these deviating tracts are the same on the cingulum and the corpus callosum, the variability is signficantly higher in the cingulum due to the relatively low number of tracts in the cingulum bundle as opposed to the corpus callosum.

The resulting atlas through the proposed algorithm is also presented as isoprobability surfaces to qualitatively inspect the resulting distributions in 3D (Figure 6-9). These spatial distributions retained very little of the tractography noise that was apparent in the individuals' tract bundles, and conformed well with the known anatomy.

Figure 6-9: Isoprobability surfaces of the spatial distributions of the fornix (in green) and the cingulum (in purple) bundles constructed from 15 subjects using the EM algorithm with tract cut operations. The surface is generated by thresholding the spatial atlas at a probability level of 10^{-4}. A few selected tracts from the corpus callosum (in black) are also drawn to highlight the spatial proximity of the three bundles. These spatial distributions retain very little of the tractography noise that is apparent in the individuals' tract bundles.

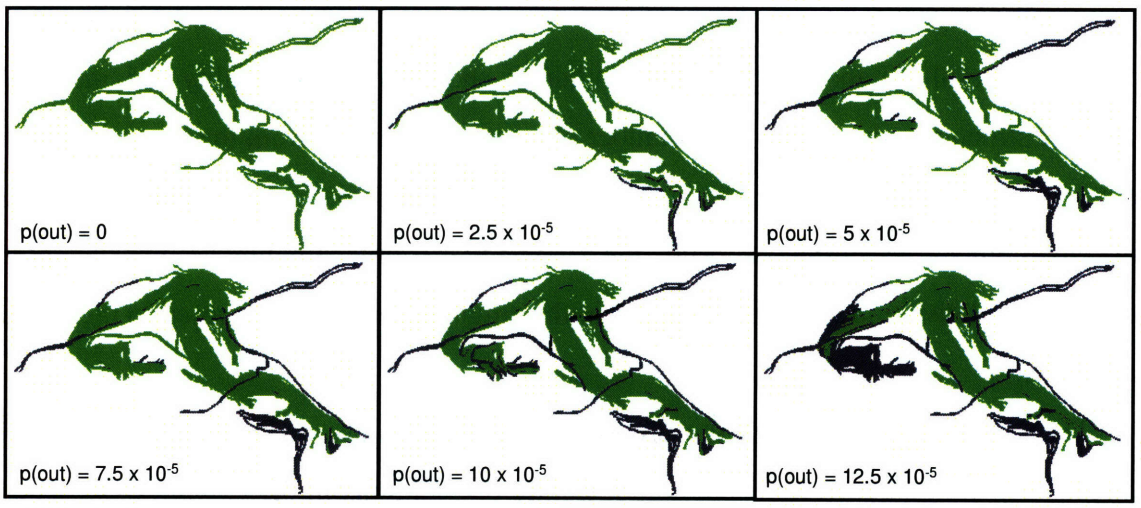

Figure 6-10: Impact of varying outlier class probability level demonstrated on the fornix of one exemplar subject. As the outlier class probability is increased, a larger percentage of the tracts from the fornix (in green) are classified as outliers (in gray).

6.5 Summary of the Chapter

In this Chapter, we extended the consistency clustering algorithm to segment and label white matter fibers while computing an atlas of fiber-bundles from a set of multi-subject diffusion weighted MR images. To demonstrate the use of this algorithm, we constructed spatial distributions of three fiber bundles in close proximity. The relative closeness of these bundles causes erroneous labeling of tracts when automatically labeled with a high dimensional atlas. However, the integrated approach we present in this paper is able to recover from these errors, resulting in a sharper atlas that is less contaminated with the mentioned artifacts.

We also demonstrated the stability of our algorithm by randomly changing the initial labels of varying proportion of the tracts. The results indicate that the solution of Consistency Clustering is stable under initialization and therefore is well suited for medical applications.

Another important aspect of any atlas is its ability to label novel subjects. We demonstrated that the tract labels are very similar whether a subject is used in the atlas building stage, or is left out at that stage and labeled afterward with the resulting atlas.

Furthermore, we presented a fully automated atlas construction pipeline, that consists of tractography, labeling with a high dimensional atlas and group-wise correction of the labels and registrations. With the introduction of a principled outlier rejection and tract cut scheme, we showed that it is also possible to retain only the "core" tract bundles, the tract segments that have their counterparts present in the group of subjects that is being analyzed. Since all of the fiber bundles are also aligned in a common coordinate system, it is then possible to define measurement axes, which are not limited to simple curves, for each of the core bundles and collect quantitative measurements for group-wise morphometry. This could prove to be a powerful statistical framework for the comparison of different sub-populations, to identify disease-related or developmental changes in the white matter anatomy.

Due to the low SNR and ambiguity of the tensor model in the fiber crossings, streamline tractography results in many errors. We have introduced a tract cut operation to identify and

reduce these artifacts. It is of future interest, however, to include the tractography into this iterative framework to make these corrections in a more integrated way. The probabilistic nature of our algorithm makes it possible for such an extension and the fact that we are using spatial and voxelized atlases makes it possible to fuse DTI based white-matter atlases with ones that were created using expert labeled structural MR images.

Chapter 7

Conclusion

In this thesis, we investigated automatic atlas construction for population-wise consistent segmentations of medical images. The atlases were constructed using an integrated registration and clustering algorithm, and evaluated by computing an anatomical atlas of fiber-bundles and deep gray matter nuclei from a population of diffusion tensor MR images (DT-MRI).

The consistency between segmentations in different subjects is ensured through using a single anatomical model for the whole population, which serves as an anatomical atlas. We experimented with both parametric and non-parametric models for the gray matter and white matter segmentation problems, each model resulting in a different kind of atlas. We constructed these atlases using unsupervised and semi-supervised approaches, which is enabled by our problem formulation for the population-wise segmentation. We refer to the integrated population-wise segmentation algorithm as Consistency Clustering since the outputs of the algorithm include population-wise consistent segmentations and correspondence between the subjects.

Consistent population-wise segmentations required development of several integrated algorithms for clustering, registration, atlas-building and outlier rejection. In this thesis we developed, implemented and evaluated these tools individually and together as a population-wise segmentation tool. We evaluated Consistency Clustering using in vivo DT-MRI datasets. We investigated the benefits of population-wise segmentation as opposed to individually segmenting subjects, as well as effects of noise and initialization on the segmentations.

Consistency Clustering provided consistent segmentations for the population, either with prior information about the expert labels or in a completely data-driven fashion. Also, through the use of labels from other subjects in the population, the algorithm was able to produce segmentations that were both qualitatively and quantitatively very similar to the expert's preference.

Furthermore, along with noisy labels generated from a High Dimensional Atlas or a single labeled subject, Consistency Clustering becomes a fully automated atlas construction pipeline for the white matter tracts. With the introduction of a principled outlier rejection and tract cut scheme, we showed that it is possible to retain only the "core" tract bundles, the tract segments that have their counterparts present in the group of subjects that is being analyzed. Since all of the fiber bundles are also aligned in a common coordinate system, it is then possible to define measurement axes, which are not limited to simple curves, for each of the core bundles and collect quantitative measurements for group-wise morphometry. This could prove to be a powerful statistical framework for the comparison of different sub-populations, to identify disease-related or developmental changes in the white matter anatomy.

Consistency Clustering can be used in two different ways to produce consistent segmentations in a population. The first way involves running the algorithm unsupervised on the population, and then assigning anatomical labels to the segmentations only on one of the subjects. The labels are then automatically transferred to the rest of the subjects since the correspondence problem is already solved at this stage. The second way involves labeling one or several subjects by hand, and then using these labeled subjects as prior information to label the rest of the population according to the expert preference.

Either way, Consistency Clustering is a powerful tool that provides fast and consistent segmentation of the deep gray matter as well as white matter tracts and has a use in a variety of applications such as in quantitative morphometry studies and pre-surgical planning. The data driven segmentation is useful in certain settings, such as when there is limited expert labeled data is available for the structure to be segmented or when expert labeled data is obtained for normal subjects but the novel subjects (that are to be segmented) are affected by a disease. In these settings, obtaining a consistent segmentation with no or little

dependence on expert labeled data, makes it possible for the expert to efficiently assign consistent anatomical labels to the whole population, and quantify population-wise impact of the disease under investigation.

Future Research Directions

Consistency Clustering requires a data model and learns the parameters of the model for the population under investigation. The accuracy of the segmentations depend on how accurately the model can represent the population. In this thesis, we used simple models for the deep gray matter and the white matter tracts, since simple models result in simple algorithms. Specifically, for the gray matter we used a Gaussian spatial model and assumed independence between the locations of the voxels. An improvement over this model includes a higher order spatial model and a Markov random field so that the neighboring voxels impact each other. We also used only the principal diffusion direction information from the rich but noisy diffusion weighted images. The principal diffusion direction may not be informative at the fiber crossing locations, which could more accurately modeled by fitting multiple tensors to the diffusion data. Similarly, we modeled white matter tracts as a bag of samples, which does not serve as a generative model for the tracts. Once again, it is possible to use higher order models that represent the data better, and these models could increase the accuracy of the results; albeit at an additional algorithmic complexity. Furthermore, in both gray matter and the white matter we used a poly-affine registration framework, which can be readily replaced by a more flexible registration framework.

In the thesis, we compared our automatic gray matter segmentation results against labels produced by a single expert. Further validation with additional experts is necessary to use Consistency Clustering in a clinical setting, such as for pre-surgical planning. In the case when additional expert labels for the thalamic nuclei proves challenging to obtain, it is possible to use other validation methods, such as tracing the white matter tracts from the labeled thalami to the cortex, for which the anatomical labels are more readily available. More immediately, Consistency Clustering can prove useful as an interactive segmentation since the framework is adequately flexible to include any amount of labeled data to be pooled in, and achieves convergence in linear time.

Bibliography

[1] D. C. Alexander, J. C. Gee, and Bajcsy R. Similarity measures for matching diffusion tensor images. In *BMVC*, 1999.

[2] D. C. Alexander, C. Pierpaoli, P. J. Basser, and J. C. Gee. Spatial transformations of diffusion tensor magnetic resonance images. *IEEE Transactions of Medical Imaging*, 20(11), November 2001.

[3] A. Banerjee, I. S. Dhillon, J. Ghosh, and Sra S. Clustering on the unit hypersphere using von Mises-Fisher distributions. *J. Mach. Learn. Res.*, 6:1345–1382, 2005.

[4] P. J. Basser and D. K. Jones. Diffusion-Tensor MRI: Theory, experimental design and data analysis - a technical review. *NMR in Biomedicine*, 15:456–467, 2002.

[5] P. J. Basser, J. Mattiello, and D. LeBihan. Estimation of the effective self-diffusion tensor from the NMR spin echo. *J Magn Reson B*, 103(3):247–254, 1994.

[6] P.J. Basser, J. Mattiello, and D. Le Bihan. MR diffusion tensor spectroscopy and imaging. *Biophys. J.*, 66:259–267, 1994.

[7] P.J. Basser, S. Pajevic, C. Pierpaoli, J. Duda, and A. Aldroubi. In vivo fiber tractography using DT–MRI data. *Magnetic Resonance in Medicine*, 44:625–632, 2000.

[8] P.J. Basser and C. Pierpaoli. Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI. *J. Magn. Reson. Ser. B*, 111:209–219, 1996.

[9] T. E. Behrens, H. Johansen-Berg, M. W. Woolrich, S. M. Smith, C. A. Wheeler-Kingshott, P. A. Boulby, G. J. Barker, E. L. Sillery, K. Sheehan, O. Ciccarelli, A. J. Thompson, J. M. Brady, and P. M. Matthews. Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. *Nat Neurosci*, 6(7):750–757, July 2003.

[10] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford Press, New York, 1997.

[11] R. Brown. A brief accouont of microscopical observations made in the months of june, july, and august 1827 on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. *Philosop. Mag.*, 4:161–173, 1828.

[12] P. T. Callaghan. *Principles of Nuclear Magnetic Resonance Microscopy*. Oxford University Press, 1993.

[13] O. Ciccarelli, G.J.M. Parker, A.T. Toosy, C.A.M. Wheeler-Kingshott, G.J. Barker, P.A. Boulby, D.H. Miller, and A.J. Thompson. From diffusion tractography to quantitative white matter tract measures: a reproducibility study. *NeuroImage*, 18:348–359, 2003.

[14] O. Ciccarelli, A. T. Toosy, G. J. M. Parker, Wheeler C. A. M. Kingshott, G. J. Barker, D. H. Miller, and A. J. Thompson. Diffusion tractography based group mapping of major white-matter pathways in the human brain. *NeuroImage*, 19(4):1545–1555, August 2003.

[15] Luis Concha, Christian Beaulieu, and Donald W. Gross. Bilateral limbic diffusion abnormalities in unilateral temporal lobe epilepsy. *Annals of Neurology*, 57(2):188 – 196, 2004.

[16] T. E. Conturo, N. F. Lori, T. S. Cull, E. Akbuda, A. Z. Snyder, J. S. Shimony, R. C. McKinstry, H. Burton, and A. E. Raichle. Tracking neuronal fiber pathways in the living human brain. *Neurobiology*, 96:10422–10427, 1999.

[17] Isabelle Corouge, Sylvain Gouttard, and Guido Gerig. Towards a shape model of white matter fiber bundles using diffusion tensor MRI. In *ISBI*, pages 344–347, 2004.

[18] C Ding and X He. Linearized cluster assignment via spectral ordering. In *International Conference on Machine Learning*, 2004.

[19] Y. Duan, X. Li, and Y. Xi. Thalamus segmentation from diffusion tensor magnetic resonance imaging. *Journal of Biomedical Imaging*, 2, 2007.

[20] A. Einstein. *Investigations on the Theory of the Brownian Movement*. Dover Publications, 1956.

[21] C. Feddern, J. Weickert, and B. Burgeth. Level-set methods for tensor-valued images. In *Proceedings of the Second IEEE Workshop on Geometric and Level SetMethods in Computer Vision*, pages 65–72, 2003.

[22] C. Gaser, I. Nenadic, B. R. Buchsbaum, E. A. Hazlett, and M. S. Buchsbaum. Ventricular enlargement in schizophrenia related to volume reduction of the thalamus, striatum, and superior temporal cortex. *Am J Psychiatry.*, 161:154–156, 2004.

[23] M. L. Giroux and et. al. Medication related changes in cerebral glucose metabolism in Parkinson's disease. In *ICFMHB*, page 237, 1998.

[24] M. L. Giroux, S. T. Grafton, J. R. Votaw, M. L. DeLong, and J. M. Hoffman. Medication related changes in cerebral glucose metabolism in Parkinson's disease. In *HBM*, page 237, 1998.

[25] H. Gray. *Grays Anatomy.* Lea and Febiger, 20th U.S. edition edition, 1918, 2000.

[26] J. Guridi, M. C. Rodriguez-Oroz, A. M. Lozano, E. Moro, A. Albanese, B. Nuttin, J. Gybels, E. Ramos, and J. A. Obeso. Targeting the basal ganglia for deep brain stimulation in parkinson's disease. *Neurology*, 55:S21–S28, 2000.

[27] J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. *Appl. Statist.*, 28:100–108, 1979.

[28] E. Heiervang, T. E. Behrens, C. E. Mackay, M. D. Robson, and H. Johansen-Berg. Between session reproducibility and between subject variability of diffusion MR and tractography measures. *NeuroImage*, 33:867–877, 2006.

[29] D. Higham and M. Kibble. A unified view of spectral clustering. *Mathematics Research Report*, pages 1–17, 2004.

[30] M. Jenkinson, P. Bannister, M. Brady, and S. Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841, 2002.

[31] L. Jonasson, X. Bresson, P. Hagmann, O. Cuisenaire, R. Meuli, and J.-P. Thiran. White matter fiber tract segmentation in DT-MRI using geometric flows. *Medical Image Analysis*, 9:223–236, 2005.

[32] L. Jonasson, P. Hagmann, C. Pollo, X. Bresson, C. R. Wilson, R. Meuli, and J.-P. Thiran. A level set method for segmentation of the thalamus and its nuclei in DT-MRI. *Signal Process.*, 87(2):309–321, 2007.

[33] L. Jonasson, P. Hagmann, C. Richero Wilson, X. Bresson, C. Pollo, R. Meuli, and J. P. Thiran. Coupled, region based level sets for segmentation of the thalamus and its subnuclei in DT-MRI. In *ISMRM*, page 731, 2005.

[34] D. K. Jones, Horsfield M. A., and Simmons A. Optimal strategies for measuring diffusion in anisotropic systems by magnetic resonance imaging. *Magn. Reson. Med.*, 42:515–525, 1999.

[35] D. K. Jones, M. A. Horsfield, and A. Simmons. Optimal strategies for measuring diffusion in anisotropic systems by magnetic resonance imaging. *MRM*, 42(3):515–525, September 1999.

[36] E.G. Jones. Making brain connections: neuroanatomy and the work of TPS powell 1923-1996. *Annual Review of Neuroscience*, 22:49–103, 1999.

[37] R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad and spectral. *J ACM*, 51(3):497–515, 2004.

[38] C. Kidwell, Chalela J., Saver J., and et al. Comparison of mri and ct for detection of acute intracerebral hemorrhage. *JAMA*, 292:1823–1830, 2004.

[39] M. Kubicki, H.-J. Park, C.-F. Westin, P. Nestor, R. Mulkern, S. E. Maier, M. Niznikiewicz, E. Connor, J. Levitt, M. Frumin, R. Kikinis, F. A. Jolesz, R. McCarley, and M. E. Shenton. DTI and MTR abnormalities in schizophrenia: Analysis of white matter integrity. *Neuroimage*, 26:1109–1118, 2005.

[40] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright. Convergence properties of the nelder-mead simplex method in low dimensions. *SIAM Journal of Optimization*, 9:112–147, 1998.

[41] D. Le Bihan, J.-F. Mangin, C. Poupon, C. A. Clark, S. Pappata, N. Molko, and H. Chabriat. Diffusion tensor imaging: Concepts and applications. *Journal of Magnetic Resonance Imaging*, 13:534–546, 2001.

[42] D. LeBihan, E. Breton, D. Lallemand, P. Grenier, E. Cabanis, and M. Laval-Jeantet. Mr imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders. *Radiology*, 161:401–407, 1986.

[43] D. LeBihan, R. Turner, and P. Douek. Is water diffusion restricted in human brain white matter? an echo-planar NMR imaging study. *Neuroreport*, 4:887–890, 1993.

[44] C. Lenglet, M. Rousson, and R. Deriche. DTI segmentation by statistical surface evolution. *IEEE Transactions in Medical Imaging*, 25:685–700, 2006.

[45] C-P. Lin, W-Y. I. Tseng, H-C. Cheng, and J-H. Chen. Validation of diffusion tensor magnetic resonance axonal fiber imaging with registered manganese-enhanced optic tracts. *NeuroImage*, 14:1035–1047, 2001.

[46] M. Maddah, A. Mewes, S. Haker, W. E. L. Grimson, and S. Warfield. Automated atlas-based clustering of white matter fiber tracts from DTMRI. In *MICCAI*, pages 188 – 195, 2005.

[47] M. Maddah, W. M. Wells, S. K. Warfield, C.-F. Westin, and W. E. L. Grimson. Probabilistic clustering and quantitative analysis of white matter fiber tracts. In *IPMI*, 2007.

[48] J. F. Mangin, C. Poupon, C. Clark, D. Le Bihan, and I. Bloch. Distortion correction and robust tensor estimation for MR diffusion imaging. *Medical Image Analysis*, 6:191–198, 2002.

[49] K. V. Mardia. *Statistical Distributions in Scientific Work*. Reidel, Dordrecht, 1975.

[50] 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. *Neuroimage*, 2(2):89–101, 1995.

[51] G. J. Mclachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley-Interscience, 2007.

[52] A. Morel, M. Magnin, and D. Jeanmonod. Multiarchitectonic and stereotactic atlas of the human thalamus. *J Comp Neurol*, 387:588–630, 1997.

[53] S. Mori, B. Crain, V. Chacko, and P. van Zijl. Three dimensional tracking ofaxonal projections in the brain by magnetic resonance imaging. *Ann. Neurol.*, 45:265–269, 1999.

[54] M. E. Moseley, Y. Cohen, Mintorovitch J., L. Chilenuittand, H. Shimizu, M. F. Kucharaczyj, M. F. Wendland, and P. R. Weinstein. Early detection of regional cerebral ischemia in cats: Comparisions of diffusion- and T2-weighted MRI and spectroscopy. *Magn. reson. Med.*, 14:330–346, 1990.

[55] J. Nolte. *The Human Brain. An Introduction to Its Functional Anatomy*. Mosby, 2002.

[56] L. O'Donnell and C.-F. Westin. White matter tract clustering and correspondence in populations. In *MICCAI*, pages 140–147, 2005.

[57] L. J. O'Donnell. *Cerebral White Matter Analysis Using Diffusion Imaging*. PhD thesis, MIT, 2006.

[58] Lauren O'Donnell and Carl-Fredrik Westin. High-dimensional white matter atlas generation and group analysis. In *MICCAI (2)*, pages 243–251, 2006.

[59] W. Ooteman and K. Cretsinger. *Thalamus Tracing Guidelines*. http://www.psychiatry.uiowa.edu/mhcrc/IPLpages/manual_tracing.htm.

[60] A Pfefferbaum, E. V. Sullivan, M Hedehus, K. O. Lim, E. Adalsteinsson, and M. Moseley. Age-related decline in brain white matter anisotropy measured with spatially corrected echo-planar diffusion tensor imaging. *Magn Reson Med.*, 44:259–268, 2000.

[61] C. Pierpaoli and P. J. Basser. Toward a quantitative assessment of diffusion anisotropy. *Magnetric Resonance in Medicine*, 36:893–906, 1996.

[62] K.M. Pohl. *Prior Information for Brain Parcellation*. PhD thesis, MIT, 2006.

[63] C. Portas, J. Goldstein, M. Shenton, H. Hokama, C. Wible, I. Fischer, R. Kikinis, R. Donnino, F. Jolesz, and R. McCarley. Volumetric evaluation of the thalamus in schizophrenic male patients using magnetic resonance imaging. *Biol. Psych.*, pages 649–659, 1998.

[64] T. G. Reese, O. Heid, R. M. Weisskoff, and V. J. Wedeen. Reduction of eddy-current-induced distortion in diffusion MRI using a twice-refocused spin echo. *MRM*, 49(1):177–182, 2003.

[65] M. Rousson, C. Lenglet, and R. Deriche. Level set and region based surface propagation for diffusion tensor MRI segmentation. In *MMBIA*, 2004.

[66] J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI*, 22(8):888–905, 2000.

[67] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *PAMI*, 22(8):888–905, 2000.

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

[69] E. O. Stejskal and J. E. Tanner. Spin diffusion measurements: Spin echoes in the presence of a time-dependent field gradient. *Journal of Chemical Physics*, 42:288–292, 1965.

[70] D. G. Taylor and M. C. Bushell. The spatial mapping of translational diffusion coefficients by the NMR imaging technique. *Phys. Med. Biol.*, 30:345–349, 1985.

[71] Dave S. Tuch. *Diffusion MRI of Complex Tissue Structure*. PhD thesis, MIT, 2002.

[72] Z Wang and B. C. Vemuri. An affine invariant tensor dissimilarity measure and its applications to tensor-valued image segmentation. In *CVPR (1)*, pages 228–233, 2004.

[73] Z. Wang and B. C. Vemuri. Tensor field segmentation using region based active contour model. In *ECCV*, pages 304–315, 2004.

[74] Z. Wang and B. C. Vemuri. DTI segmentation using an information theoretic tensor dissimilarity measure. *IEEE Transactions in Medical Imaging*, 24:1267–1277, 2005.

[75] S.K. Warfield, K.H. Zou, and W.M. Wells. Simultaneous truth and performancelevel estimation (STAPLE): An algorithm for the validation of imagesegmentation. *IEEE Transactions in Medical Imaging*, 2004.

[76] V. Wedeen, T. Reese, D. Tuch, M. Wiegel, J.-G. Dou, R. Weiskoff, , and D. Chessler. Mapping fiber orientation spectra in cerebral white matter with Fourier transformdiffusion MRI. In *ISMRM*, page 82, 2000.

[77] Yair Weiss. Segmentation using eigenvectors: A unifying view. In *ICCV*, page 975, 1999.

[78] C.-F. Westin, S.E. Maier, B. Khidhir, P. Everett, F.A. Jolesz, and R. Kikinis. Image Processing for Diffusion Tensor Magnetic Resonance Imaging. In *miccai*, Lecture Notes in Computer Science, pages 441–452, September 1999.

[79] C.-F. Westin, S.E. Maier, H. Mamata, A. Nabavi, F.A. Jolesz, and R. Kikinis. Processing and visualization of diffusion tensor MRI. *Medical Image Analysis*, 6(2):93–108, 2002.

[80] C.-F. Westin, S. Peled, H. Gudbjartsson, R. Kikinis, and F.A Jolesz. Geometrical diffusion measures for MRI from tensor basis analysis. In *ISMRM '97*, Vancouver, Canada, April 1997.

[81] M.R. Wiegell, D.S. Tuch, H.W.B. Larson, and V.J. Wedeen. Automatic segmentation of thalamic nuclei from diffusion tensor magnetic resonance imaging. *Neuroimage*, 19:391–402, 2003.

[82] Terence H. Williams, Nedzad Gluhbegovic, and Jean Y. Jew. *The Human Brain: Dissections of the Real Brain*. Virtual Hospital, 1997.

[83] Dongrong Xu, Susumu Mori, Meiyappan Solaiyappan, Peter C. M. van Zijl, and Christos Davatzikos. A framework for callosal fiber distribution analysis. *NeuroImage*, 17:1131–1143, 2002.

[84] J. L. Young and et. al. *SEER's Training Site*. http://training.seer.cancer.gov/.

[85] L. Zhukov, K. Museth, D. Breen, R. Whitakert, and A. H. Barr. Level set modeling and segmentation of DT-MRI brain data. *Journal of Electronic Imaging*, 12:125–133, 2003.

[86] U. Ziyan, D. Tuch, and C.-F. Westin. Segmentation of thalamic nuclei from DTI using spectral clustering. In *MICCAI*, pages 807–814, 2006.

[87] L. Zollei, E. Learned-Miller, W. E. L. Grimson, and W. M. Wells III. Efficient population registration of 3D data. In *ICCV, Computer Vision for Biomedical Image Applications*, 2005.

[88] K. Zou, S. Warfield, A. Bharatha, C. Tempany, M. Kaus, S. Haker, W. Wells, F. Jolesz, and R. Kikinis. Statistical validation of image segmentation quality based on a spatial overlap index. *Acad Radiol*, 11:178–89, 2004.