Sequence-Aware and Advanced Biomarker Calculation Improves Statistical Inference in Image Processing of Parkinson's Disease

By

Andrew John Plassard

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University in

partial fulfillment of the requirements for the

degree of

DOCTOR OF PHILOSOPHY

in

Computer Science

December 16, 2017

Nashville, Tennessee

Approved:

Bennett A. Landman, Ph.D.

Benoit M. Dawant, Ph.D.

Daniel O. Claassen, M.D.

Maureen K. McHugo, Ph.D.

Aniruddha S. Gokhale, Ph.D.

Table of Contents

Page

List of Tables ..... **iv**
List of Figures ..... **v**

Chapter

I INTRODUCTION 1
- **1. Parkinson's disease** **2**
- **2. Segmentation Theory** **8**
- **3. Informatics** **15**
- **4. Dissertation Focus** **17**
- **5. Narrative** **21**

II MULTI-PROTOCOL, MULTI-ATLAS STATISTICAL FUSION: THEORY AND APPLICATION 23
- **1. Introduction** **23**
- **2. Theory** **25**
- **3. Methods and Results** **31**
- **4. Discussion and Conclusion** **35**

III SYNTHETIC ATLASES IMPROVE SEGMENTATION CONSISTENCY BETWEEN T1-WEIGHTED IMAGING SEQUENCES 38
- **1. Introduction** **38**
- **2. Theory** **40**
- **3. Methods** **45**
- **4. Results** **48**
- **5. Discussion and Conclusions** **50**

IV AUTOMATED, OPEN-SOURCE SEGMENTATION OF THE HIPPOCAMPUS AND AMYGDALA WITH THE OPEN VANDERBILT ARCHIVE OF THE TEMPORAL LOBE . 52
- **1. Introduction** **52**
- **2. Methods** **54**
- **3. Results** **60**
- **4. Discussion** **64**

V IMPROVING CEREBELLAR SEGMENTATION WITH STATISTICAL FUSION 67
- **1. Introduction** **67**
- **2. Methods** **69**
- **3. Results** **75**
- **4. Discussion** **76**

VI MULTI-MODAL IMAGING WITH SPECIALIZED SEQUENCES IMPROVES ACCURACY OF THE AUTOMATED SUB-CORTICAL GREY MATTER SEGMENTATION 78
- **1. Introduction** **78**
- **2. Methods** **79**
- **3. Results** **83**
- **4. Discussion** **85**

VII IMPROVING VARIANCE ESTIMATION IN MULTI-ATLAS SEGMENTATION TO ACCURATELY CHARACTERIZE THE MARGINAL UTILITY OF ADDITIONAL ATLASES 87
1. Introduction 87
2. Theory 91
3. Methods 94
4. Results 98
5. Discussion 100

VIII CONCLUSIONS 102
1. Summary 102
2. Segmentation with Multiple Label Sets 103
3. Segmentation with Multiple Imaging Sequences 103
4. Segmentation of Specialized Anatomy 104
5. Estimation of Variance in Multi-Atlas Segmentation 104
6. Summary of Contributions 105
7. Impact on PD 106

References 107

List of Tables

Table III-1 Summary of average segmentation volumes for the three acquired scans on seven subjects. Average volumes for the two segmentation techniques, standard segmentation and synthetic segmentation, are shown. 44

Table IV-1 Atlas demographics 55

List of Figures

Figure II-1 Modeling rater performance with the STAPLE, MS-STAPLE, and SMS-STAPLE algorithms. A target (A) image can be labeled using atlases of multiple protocols. In STAPLE (B) only atlases with a 1:1 correspondence can be used in segmentation, in MS-STAPLE (C) atlases of multiple protocols can be used to jointly segment *all* of the protocols, and in SMS-STAPLE (D) atlases of multiple protocols can be used to jointly segment *one* target protocol. 24

Figure II-2 Simulation results based on algorithm and number of atlases used. A set of simulated truth models (A) were generated to model a relationship with inter-protocol spatial dependence. Simulated observations with increasing numbers of observations (B) were fused (C and D). For STAPLE v. MS-STAPLE and SMS- v. MS-STAPLE: * $p < 10^{-5}$. 32

Figure II-3 Results from empirical deep brain segmentation experiment. Both SMS-STAPLE and SMS-Non-Local STAPLE show statistically significant improvement (*$p < 0.05$, ** $p < 0.01$) for many of the segmentation tasks particularly when few atlases of the target class are available. † indicates segmentation task shown in Figure 4. 34

Figure II-4 Qualitative results comparing eight segmentation algorithms (identified by † in Figure 3). Contour lines indicate manual truth target labels (shown lower left)..... 36

Figure III-1 (A) Absolute and relative MPRAGE signals vary non-linearly based on the sequence parameters. (B) These changes introduce visually apparent boundary shifts when the same structures are compared across different imaging sequences. 39

Figure III-2 Atlases with quantitative T1-relaxation and proton density maps are used to generate synthetic atlases from sequence parameters (i.e., inversion time, echo time, repetition time). The synthesized atlases are used in the standard multi-atlas segmentation procedure. 40

Figure III-3 Volumetric segmentation results for the standard and synthetic multi-atlas segmentation approaches on seven acquired subjects. For most of the major tissue types, there was a significant reduction in variance using the synthetic segmentation compared with the standard (*, $p<0.05$). Cerebrospinal fluid did not show a significant improvement using either method..... 43

Figure III-4 Qualitative surface changes comparing the standard and synthetic multi-atlas segmentations for one randomly selected subject within ABIDE. Surface distance is measured as the distance between the surface of standard multi-atlas segmentation and the synthetic multi-atlas segmentation. Negative changes corresponds to cases where the synthetic segmentation's surface was within the standard. Synthetic segmentation tends to pull the grey matter/CSF boundary inward whereas the synthetic segmentation extrudes the white matter at similar sulci peaks. The subcortical changes are less consistent. 47

Figure III-5 Comparison of trends in acquired data compared with synthetic. The trend lines and points for one subject of each are shown here, normalized to a mean of zero. Six tissue types are presented here and cerebral white matter, cortical grey natter, subcortical grey matter, cerebellar grey matter, and cerebellar white matter showed similar trends between the synthetic and acquired images. Similar trends are present across the whole population and are statistically significantly similar between the acquired and synthetic data. 46

Figure III-6 In the leave-site-out classification (A) of autism versus control, the mean AUC was significantly higher using the synthetic segmentation volumes compared with the standard segmentations. Consistent trends in percent volume difference between the standard and synthetic segmentations were seen for these subjects (B). Site was significantly correlated with percent volume difference ($p<0.01$, Pearson's correlation) whereas age was not..... 49

Figure IV-1: Quantitative segmentation results for the whole hippocampus and amygdala. OVAL outperforms all other segmentation techniques in terms of DSC for the left hippocampus in both raters, the right hippocampus in rater 2, and the left and right amygdala ($p<0.05$, *). OVAL outperforms all other techniques for the right hippocampus of rater 1 except human reproducibility, which performs comparably statistically. Human reproducibility outperforms all other techniques in MSD for the left and right hippocampus for rater 1 ($p<0.05$, *). OVAL and OVAL with 30 atlases outperform all other automated techniques for those structures. OVAL, OVAL with 30 atlases, and human reproducibility perform statistically comparable for the left and right hippocampus for rater 2 and outperform all other techniques ($p<0.05$, *). OVAL and FSL FIRST perform statistically similarly for the left amygdala and outperform all other segmentation approaches ($p<0.05$, *). OVAL, OVAL with 30 atlases, and FSL FIRST perform statistically similarly for the right amygdala and outperform all other segmentation approaches ($p<0.05$, *). 56

Figure IV-2 Quantitative segmentation results for the whole hippocampus head, body, and tail. OVAL outperformed OVAL with 30 atlases and human reproducibility on the right head and body in Dice Similarity Coefficient ($p<0.05$; *). No technique showed significant improvement on the right or left tail. Human reproducibility outperformed OVAL and OVAL with 30 atlases ($p<0.05$, *), though OVAL outperformed OVAL with 30 atlases. In mean

surface distance, OVAL and human reproducibility outperformed OVAL with 30 atlases for the right head, OVAL outperformed all other techniques for the right body, no technique outperformed any other for the right and left tail, and human reproducibility outperformed the other techniques for the left and right head ($p<0.05$, *). 59

Figure IV-3 Median qualitative segmentation results for the whole hippocampus and amygdala; red represents the estimated segmentation and green is the truth. FSL FIRST, BrainCOLOR, and FreeSurfer all showed large surface distances up to 4mm for both the hippocampus and amygdala. OVAL and OVAL with 30 atlases were typically within 1mm distance on the hippocampus, though OVAL produced more consistent results than OVAL with 30 atlases. On the amygdala, OVAL and OVAL with 30 atlases captured the overall contour of the amygdala, but were not able to accurately localize the borders since they are defined by anatomical landmarks. 61

Figure IV-4 Median qualitative segmentation results for the hippocampus head, body, and tail. Green represents the true segmentation and red represents the estimate. Human reproducibility defined a different point for the head/body split and rater 2 under-segmented the tail of the hippocampus and the tip of the head compared with rater 1. OVAL with 30 atlases produced more local errors than OVAL. Images were rotated along the axis of the hippocampus, gaps between the head, body, and tail are exaggerated for visualization. 63

Figure V-1: Axial, coronal, and sagittal segmentation results for a healthy (A) subject and a patient with severe cerebellar ataxia (B). Note the easily differentiable lobules in the patient whereas the differentiation of the lobules is lost to the resolution of the imaging in the healthy subject. 68

Figure V-2 Summarized segmentation results for the Anura and AT datasets. Non-Local SIMPLE outperformed all other techniques on the Anura dataset (A). On the AT dataset Non-Locally Weighted Vote significantly outperformed all other techniques, but Non-Local SIMPLE still outperformed the previously gold-standard technique of Yang et al (A).Qualitatively, Non-Locally Weighted Vote seemed to oversegment the lobules whereas Non-Local SIMPLE tended to undersegment. The results of Yang et al visually produced results more consistent with the anatomic boundaries but had more internal boundary shifts than either Non-Locally Weighted Vote or Non-Local SIMPLE. 70

Figure V-3 Quantitative segmentation results for the Anura dataset. Non-Local SIMPLE shows either significant improvements over other algorithms or comparable results to other algorithms for all labels. 72

Figure V-4 Quantitative segmentation results for the Ataxia dataset. No algorithm shows significant improvement across all labels but Non-Locally Weighted Vote provides both consistent and accurate results across most labels. 73

Figure V-5 Qualitative segmentation results from the median Ataxia subject. Non-Locally Weighted Vote tends to slightly over-segment regions of interest while Non-Local SIMPLE tends to under-segment regions. The adaptation of Yang et. al appears to generate a segmentation more consistent with anatomic boundaries but can produce severe missegmentations as seen in the sagittal view. Other algorithms are not shown since they infrequently outperformed the algorithms shown here. 75

Figure VI-1 Segmentation results for structures in the dienchephalon. Quantitative segmentation results are shown in (A). For the left SN, multi-modal segmentation with T1 and FGATIR

outperformed other approaches (*; p<0.05; Wilcoxon sign-rank test). For the right SN no segmentation approach outperformed other approaches. For the left STN, multi-modal segmentation with T1 and FGATIR outperformed other approaches (*; p<0.05; Wilcoxon sign-rank test). For the right STN no segmentation approach outperformed other approaches. In (B), surface distances between the true and estimated segmentations for the left SN are shown for the six proposed segmentation approaches. 81

Figure VI-2 Segmentation results for structures in the globus pallidus. Quantitative segmentation results are shown in (A). For the left GPE, multi-modal segmentation with T1 and FGATIR with double atlases outperformed other approaches (*; p<0.05; Wilcoxon sign-rank test). For the right GPE segmentation with FGATIR outperformed other approaches (*; p<0.05; Wilcoxon sign-rank test). For the left GPI, multi-modal segmentation with T1 and FGATIR with doubled atlases and segmentation with FGATIR with doubled atlases outperformed other approaches but were not distinguishable amongst each other (*; p<0.05; Wilcoxon sign-rank test). For the right GPI no segmentation approach outperformed other approaches. In (B), surface distances between the true and estimated segmentations for the left GPI are shown for the six proposed segmentation approaches. 82

Figure VI-3 Segmentation results for the putamen and thalamus. Quantitative segmentation results are shown in (A). For the left putamen, multi-modal segmentation with T1 and FGATIR with double atlases outperformed other approaches (*; p<0.05; Wilcoxon sign-rank test). For the right putamen segmentation with FGATIR with doubled atlases outperformed other approaches (*; p<0.05; Wilcoxon sign-rank test). For the left thalamus, no segmentation approach outperformed other approaches For the right thalamus, no segmentation approach outperformed other approaches. In (B), surface distances between the true and estimated

segmentations for the left putamen are shown for the six proposed segmentation approaches. 84

Figure VII-1 Example segmentations of the hippocampus using 20 atlases (A). Estimate 1 and Estimate 2 use unique populations of atlases. The surface estimates show significantly different estimates at several boundary locations (identified by the green arrows). Monte Carlo estimates of confidence intervals become invalid especially as the number of atlases in the sample approaches the number of atlases available (B). In cases where the total population size is 100 (red curves) and 200 (blue curves), the variance estimated by Monte Carlo approached 0 as the number of atlases used approached the total pool size..... 88

Figure VII-2 Variance estimation results from Monte Carlo sampling. In (A) each curve represents a pool of atlases of increasing size. The number of atlases are sampled from the pool in increasing amounts. These results are then interpolated with Expected Percent Similarity (B) to identify a consistent pattern of decline of variance with respect the Expected Percent Similarity. (C) and (D) show the patterns of (A) and (B) with all sizes of pools sampled... 90

Figure VII-3 Example fits of the Expected Percent Similarity to the variance estimate (A) for increasing numbers of atlases. Variance values are scaled to their log for visualization purposes. (B) the final fit results for the estimated variance from the proposed algorithm, compared with the pooling approach, standard Monte Carlo, and bootstrapping approaches. The estimated approach follows a similar trend to the bootstrapping and pooling based estimates. On the other hand, the Monte Carlo approach deviated from other approaches in particular as the number of atlases considered approached the total available..... 93

Figure VII-4 Re-fit variance results for the Monte Carlo estimation approach (A). These variance estimations do not converge as the standard Monte Carlo variance estimation did. These results show that variance is still decreasing as the number of atlases increases, and the average is also increasing. Comparing the final distribution with 180 atlases to the distributions with fewer atlases, we show that 180 atlases outperforms atlas counts up to 160 (B). Furthermore, when comparing the results for identifying the proper number of atlases from the population size of 150, the estimated variance approach identified that at least 150 atlases were needed whereas the Monte-Carlo variance approach identified 136 as the optimal number of atlases (C). 96

Figure VII-5 Percent of the total voxels likely to change between two unique draws of atlases from a population. The x-axis shows increasing numbers of atlases in the segmentation. Each color represents one of the subjects on which the experiment was performed. The filled in lines represent the estimates from the proposed approach, whereas the dotted lines represent the Monte Carlo estimates. We can see that as more atlases are added, a smaller portion of voxels are likely to change between draws from the atlas population. The Monte Carlo estimates tend to underestimate the number of voxels likely to change. The proposed method shows that, though the rate of change is decreasing of the number of voxels likely to change between two random draws, that there is still added value by adding additional atlases. 98

Chapter I

Introduction

Improved segmentation of magnetic resonance imaging is necessary to provide quantitative and anatomical information for current and future analytic and radiological understanding of Parkinson's disease and progression. Image Segmentation is a common task, accomplished by a family of approaches, for calculation of volumetric and structural biomarkers, endophenotypes useful for univariate or multivariate prediction of disease state and progression, in medical images. In Parkinson's there is a focus on understanding subcortical grey matter, in particular as localization for deep brain stimulation surgery.

Improvements in imaging of disease states has rapidly outpaced the availability of image processing approaches. Best practices in study design have not kept pace with the new modalities and approaches. For example, a wealth of data is available from studies like the Autism Brain Imaging Data Exchange (ABIDE) where different scanning sequences were used at each site, potentially biasing calculations based on where each subject was scanned. In order to study a disease like Parkinson's disease, a population of both healthy and diseased patients scanned under an identical protocol are needed to make proper inference. Currently, datasets acquired are not amenable to standard segmentation approaches and thus make inter-study and retrospective analyses challenging due to biases induced by target imaging sequence variation and atlas populations.

Image processing is an important step in biomarker acquisition from medical imaging data. In order to improve our statistical power and derive as much information as possible from our imaging approaches, the approaches developed must be aware of the biases and assumptions

present in the imaging. Image processing approaches which account for the sequence that the target sequence, and are invariant to other biases in the target data, will have improved statistical power due to less variance form factors external to the factors in question.

The remainder of this section proceeds as follows. First we discuss the anatomic phenotypes related to Parkinson's disease. Second, we discuss the theory behind which a segmentation is performed. Third we discuss the informatic challenges behind image processing at Vanderbilt University. Finally, we describe the contributions of this work.

1. Parkinson's disease

Parkinson's Disease (PD) is a neurodegenerative disorder primarily effecting the central nervous system (CNS) [1]. In the early stages of PD, patients suffer from motor instabilities including tremors, rigidity, and gait instability [2]. In later stages. PD patients can develop severe neuropsychiatric symptoms; PD patients have over twice the likelihood of dementia compared with the general public [3], impaired executive function [4], and various mood disorders [2]. In 2013, PD effected 53 million people and resulted in 103,000 deaths, making it the second most prevalent neurodegenerative disease [5].

There is no diagnostic test or exact criteria for diagnosis of PD. PD diagnosis is based on physician review of a patient's medical history and neurological examination [6, 7]. Often times, diagnosis is reinforced with decreased motor symptoms after receiving Levadopa (L-DOPA). Several clinical organizations provide diagnostic criteria to diagnose PD, but these criteria often require 5-10 years of PD symptoms [2]. Diagnosis can only be confirmed by autopsy and the presence of alpha-synuclein build-up in the midbrain [8, 9]. These alpha-synuclein build-ups, known as Lewy Bodies, are the distinguishing characteristic of PD, but their pathogenesis and role

are not well understood [10, 11]. Even with the diagnostic criteria, the success rate of PD diagnosis is 70-95% [2].

1.1. Neurophysiology of Parkinson's Disease

PD is primarily described as a disease of the basal ganglia, a series of structures in the mid-brain consisting of the striatum, globus pallidus internal and extrenal, subthalamic nucleus, and substantia nigra [12]. There has also been significant interest in the limbic system [13], thalamus [6], and cerebellum [14], all of which are connective neighbors to the basal ganglia.

1.1.1. Basal Ganglia

The basal ganglia is a collection of subcortical nuclei located bilaterally in the cerebrum. The basal ganglia is involved in cognitive processing, motor control, and procedural learning [15-17]. The striatum is the receives glutamatergic and dopaminergic inputs and serves as the primary input to the basal ganglia [18]. The external globus pallidus (GPE) receives dopaminergic signal from the subthalamic nucleus and has signaling neurons projecting to other parts of the basal ganglia. The internal globus pallidus (GPI) is an output neuron of the basal ganglia, receiving signals from the subthalamic nucleus and sending signals to the thalamus [19, 20]. The subthalamic nucleus (STN) is a transmission nucleus in the basal ganglia, involved in action selection and reward control [21]. The substantia nigra is the second output, sending signals from the striatum to various parts of the brain [22].

In PD the basal ganglia is effected in several ways. The dorsal rostral portion of the striatum has been shown to have a severe reduction of dopamine [23]. The external globus pallidus has shown a unique, rhythmic spiking in its neurons [24]. The internal globus pallidus has shown

a significant decrease in dopaminergic discharge [25]. In a monkey model of PD the subthalamic nucleus showed an increase in tremor neuronal activity [26]. The substantia nigra showed significant regional atrophy, particularly in the lateral ventral tier, which differs from the typical pattern of aging [27].

1.1.2. Thalamus

Located bilaterally in the diencephalon, the thalamus is a centralized relay point between the cortex and other portions of the brain [28, 29]. Each lateralized portion of the thalamus is subdivided into specialized nuclei. Divided by the internal medullary and two lamina, the medial, anterior, and lateral nuclei groups constitute the major regions of the thalamus [30]. The medial nuclei relays signals from the limbic system, in particular the amygdala, the prefrontal cortex, and the olfactory cortex [31, 32]. The medial nuclei are hypothesized to be involved in pain processing and memory [33]. The anterior thalamic nuclei receive input neurons from the hypothalamus and the subiculum through the fornix and relay to the cingulate gyrus [34]. The anterior nuclei are involved in spatial localization and navigation [35]. These nuclei are also considered to be a part of the limbic system. The lateral nuclei group is the largest of the thalamus and receives inputs from the visual cortex, internally from other thalamic nuclei, the posterior parietal cortex, and the cingulate [28, 36]. The lateral nuclei are involved in visual attention and attention deficit [37].

Due to the size and lack of contrast on MR imaging, little is known about the importance of the nuclei groups in PD. Most studies considering the role of the thalamic nuclei in PD use post-mortem studies of model organisms [38, 39]. One of the major findings of these studies is a decrease in non-dopaminergic cells in the inter-laminar thalamic nuclei [40]. Other studies have shown that the pedunculopontine nuclei, in particular its feedback loop with the basal ganglia,

spinal cord, and limbic system, are related to motor instability and gait difficulty [41]. In humans, stimulation of the ventral intermediate thalamic nucleus has been shown to reduce or eliminate tremors in both PD and a related disorder, essential tremor (ET) [39].

1.1.3. Cerebellum

The cerebellum, or "little brain", is a sub-structure of the human brain with known implications in motor coordination and degenerative disorders with hypothesized involvement in cognitive function and emotional regulation [42, 43]. Located beneath the cerebrum in the posterior fossa, the cerebellum is a secondary feedback loop for the spinal cord and basal ganglia [44]. Anatomically, the cerebellum is divided into a left and right hemisphere, connected by a midline vermal layer [42]. The cerebellum has a cortical layer, similar to the cerebrum, with white matter beneath it, connecting the cerebellum to the remainder of the brain and spinal cord. Within the white cerebellar white matter, four bilateral deep nuclei, the dentate, emboliform, globose, and fastigii nuclie, receive GABAergic signals from the cerebellar cortex and originate most of the output fibers from the cerebellum [45]. The hemispheres of the cerebellum are sub-divided into three major lobes, the anterior, posterior, and flocculonodular lobe, based on major fissures in the cerebellum and ten lobules based on the folds of the cerebellum [46].

In PD, the cerebellum has been implicated in many of the primary phenotypes, but has not been researched to the same degree as the basal ganglia [47]. Recently, functional MRI has shown that tremors may originate from a functional network between the motor cortex, basal ganglia, and cerebellar vermis [24]. Results from a PET study show a correlation between balance and gait instability in PD and acetylcholinesterase activity in the mid-brain and cerebellum [48]. Grey matter volume in the cerebellum was implicated in impaired cognition in non-demented PD [49].

In general, these results are presented for the full cerebellum, but the particular lobes and lobules have not been particularly implicated in many studies, though given their somatotopic nature, it would naturally conclude that there should be a relationship between particular lobules and phenotypes of PD.

1.1.4. Limbic System

The limbic system is a set of brain structures involved in memory and emotional response, including cortical and sub-cortical grey matter structures, and portions of the diencephalon [50, 51]. The hypothalamus, one of the central structures of the limbic system, is located in the ventral part of the diencephalon [52, 53]. The hypothalamus contains several small nuclei which maintain numerous regulatory and metabolic systems, and is connected to the brainstem, amygdala, septum and several regions of the brain. The hippocampus, known famously for its role in memory and diseases like Alzheimer's, is a cortical structure located in the temporal lobe is involved in spatial memory and sleep [54]. Though little is known about its role in emotion and other common limbic system functions, the hippocampus is connected to the hypothalamic mammillary body, the anterior portion of the thalamus, and the amygdala [55]. The amygdala consists of a series of interconnected nuclei involved in emotional response and decision making [56]. Interestingly, the subdivisions of the amygdala are split between the limbic system and basal ganglia, making it a particularly important structure as it is hypothesized to be involved both in motor and emotional processing [12]. The nucleus accumbens is a structure in the basal forebrain, involved in motivation, reward, and fear processing, amongst other functions [57, 58]. Anatomically, the nucleus accumbens is broken into a core and shell surrounding the core, where each portion has unique neural connections. The core has neural connections with the GPI and substania nigra,

making it an important connection linking the limbic system and basal ganglia. The shell is an extension of the amygdala and involved in similar reward processing through connections to hypothalamus and amygdala.

Some research has been done on the role of the limbic system in PD, but there has been less work in comparison to the basal ganglia and thalamus. Lewy body buildup occurs in all of the sub-nuclei of the hypothalamus [59] and a significant decrease in dopaminergic activity in the hypothalamus has been related to obesity and weight loss as PD risk factors [60]. A post-mortem analysis of hippocampal dopamine and dopamine metabolites showed a significant relationship with L-DOPA dosage before death [61] which is significant because dopamine is necessary for normal memory processing in the hippocampus [62]. The amygdala has shown a particular lesioning pattern in PD which destroys the nuclear grey matter and may destroy neural connections [63]. Interestingly amygdala volume and amygdalar neuronal volume are significantly associated with PD disease status, but not correlated with disease progression [64]. The nucleus accumbens is an important processing step in dopamine uptake on L-DOPA for reversal learning tasks [65].

1.2. Treatment of Parkinson's Disease

Due to a lack of understanding of PD's mechanism of action, progression, and inception, there is no cure or treatment plan suitable for all patients [8]. Typical treatment begins with dopamine precursors, like L-DOPA, which are used to increase the dopamine content in the brain [66]. As L-DOPA treatment slows, many patients will be prescribed medications, paired with L-DOPA, targeted at increasing the biosynthesis of dopamine, but these treatments are very dependent on the patient and prescribing physician, thus their typical course is less understood [65, 66]. After treatment with medication has proven ineffective, deep brain stimulation (DBS) surgery

is used to reduce motor symptoms, but these surgeries have shown little effect on non-motor symptoms of PD [67]. Beyond medication and surgery, diet and rehabilitation have shown some effectiveness in treatment of PD, but in general are not part of standard treatment progression [68].

1.2.1. Medication for treatment of Parkinson's disease

Treatment of PD typically begins with L-DOPA, a precursor molecule to dopamine, which is a precursor to dopamine and can cross the blood-brain barrier whereas dopamine cannot. Since there is no diagnostic criteria for PD, a positive response to L-DOPA administration is typically confirmation of the diagnosis. L-DOPA mitigates motor symptoms by being converted to dopamine and increasing dopamine levels in the substantia nigra. For several reasons, treatment of PD with L-DOPA is limited and in many cases other medications are paired with L-DOPA to increase its effectiveness. Over long periods of administration of L-DOPA patients can develop involuntary movements related to L-DOPA and varied responses to the medication [65, 66].

1.2.2. Deep Brain Stimulation Surgery

When medication does not treat the motor phenotypes of PD, deep brain stimulation (DBS) surgery is an alternative to standard treatments. In DBS surgery, a neurosurgeon implants bi-lateral electrodes into the patient's brain. In PD the typical anatomic targets are the GPI and STN. After completion of the surgery, the electrodes are activated to stimulate neural activity in the motor tracts and alleviate the motor symptoms of PD, though the surgery does not always restore normal motor function.

2. Segmentation Theory

Segmentation is an important task in medical imaging. In segmentation, a target image or set of images is input into an algorithm with the goal of identifying a structure, or structures, of interest. Two broad classes of segmentation algorithm exist. The first class of algorithms is model-based segmentation algorithms, where an underlying understanding of the anatomic and physiological process is used to identify structures of interest [69]. These techniques are commonly used when there is a clear anatomic boundary between the structures of interest such as grey and white matter in T1-weighted brain images. The second class of algorithms is atlas-based segmentation techniques. In atlas-based techniques, one or more atlases, or example images with expertly delineated labels are non-rigidly registered to the target images [70, 71]. These registered atlases are then joined together to create a consensus representation of the target image's segmentation. There also exist hybrid algorithms which utilize model-based techniques in conjunction with atlases to improve the segmentation accuracy. Many gold-standard segmentation approaches use a hybrid approach where models are used in several contexts [70, 72].

2.1. Model-Based Segmentation Techniques

In model-based segmentation approaches, underlying physiological principles are leveraged to identify the structures of interest. One of the classical tasks for model-based segmentation is delineation of gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) in T1-weighted brain images. On a T1-weighted brain MRI, WM appears hyper intense and at a higher intensity than the background, CSF, or GM. CSF appears, on T1-weighted MRI, at a low intensity value and at a similar intensity value to the background. On T1-weighted MRI, GM is at an intensity spectrum between CSF and WM [73]. The intensity distributions from GM, WM, and CSF are largely non-overlapping so an algorithm like K-means or expectation maximization

(EM) with Gaussian distributions can identify the three distributions [74, 75]. The final class decision for each voxel from the K-means or Gaussian distributions can be used as the tissue classification into GM, WM, or CSF.

Markov Random Fields (MRFs) can be incorporated into segmentation algorithms to incorporate spatial information. An MRF controls the spatial correlations between nearby voxels [76]. For instance, an MRF can be used with k-means to control the probability of changing tissue types of neighboring voxels, thus incorporating spatial information into the k-means framework [77]. MRFs can also be useful in estimating gain fields in MR images and in several other contexts [78].

Deformable models, such as active shape models and active contour models, use the image as a volume and fit a curve based on the image intensities [79, 80]. This curve can then be used to divide the image into tissue classes. Active shape models fit a model of the shape of a structure of interest to a target by iteratively conforming a shape model to the target and re-assessing the point correspondence [81]. Active contour models a deformable spline to detect boundaries in an image based on image intensity [80]. Both of these models require a decent initialization of the model with respect to the target and the models are non-convex, thus results are not necessarily optimal.

Segmentation techniques are not limited to structural MR and CT images. Diffusion tensor imaging (DTI) is commonly used to model tractography within the WM. Fiber tracking uses regions of interest within the brain to identify the size and scope of WM pathways of interest [82]. These pathways can be used to identify the tract which connects given structures. For instance, TRACULA [83] utilizes the spatial relationship of well-known fiber tracts where a prior is used to constrain the paths between structures of interest.

2.2. Atlas-Based Segmentation Techniques

An atlas is an image which, commonly, has the structures of interest manually identified by experts [84]. Atlases are useful in delineating structures of interest where the prior knowledge of an expert is helpful in identifying structural boundaries that would be challenging to identify without larger contextual knowledge. For instance on standard T1-weighted images, due to the nature of white matter connections within the thalamus there is low contrast between it and the white matter [85]. A trained expert can reproducibly identify the boundaries and thus create an atlas of the thalamus.

Typically, atlas-based segmentation approaches start by non-rigidly registering one or more atlases to the target image [71]. In non-rigid registration one image volume, herein the moving image, is first affinely aligned to the target image. Affine registration is a 9 degree of freedom where the moving image can be rotated, translated, scaled, or sheered to best align with the target image [86]. For all works presented here the NiftyReg affine registration is used because it provides both an efficient and accurate affine registration [86]. After affine registration, the moving image is then non-rigidly registered to the target image. In non-rigid registration the moving image is elastically, locally warped to the target image [87]. There are several paradigms for non-rigid registration that have been previously reviewed for brain imaging. For all works presented here the Advanced Normalization Tool (ANTs) Symmetric Normalization (SyN) algorithm is used [88]. After the affine and non-rigid registrations, the labels from the atlas are deformed to the target image space following the affine and non-rigid deformations.

If only one atlas is available, the labels transferred to the target image provide an estimate of the labels for the target [71]. If multiple atlases are available, then labels from the atlases can

be combined, or "fused", as an ensemble of learners to produce a set of labels more consistent with the target image's labels than any individual label set [89]. Several algorithms exist to produce a consensus representation from a set of atlases, a process herein known as label fusion.

The simplest label fusion approach is majority vote. In majority vote, each atlas is given an equal weight. The label decision at each location is the mode label amongst the registered atlases [84]. Weighted voting algorithms determine a weight for each atlas either globally or locally [90]. These weights are commonly determined based on global or local image similarity respectively [84]. The weighted mode label at each voxel is selected as the label decision. Statistical label fusion techniques, such as Simultaneous Truth and Performance Level Estimation (STAPLE), compute a confusion matrix for each atlas [91]. A given confusion matrix consists of a matrix where entry i, j corresponds to the probability that rater observes label i when the actual label is j. This confusion matrix is calculated using the EM algorithm where at each iteration the true label probabilities are estimated at each location and then the confusion matrices are re-estimated based on the label probabilities. This process is iterated until convergence. The STAPLE framework has expanded to incorporate spatially varying confusion matrices [92], hierarchical performance [93], and a number of other improvements.

Non-local correspondence assumes that registration is not perfect and performs a local search at each location for each atlas to determine if a better correspondence can be reached [94]. For a given location and atlas, non-local correspondence searches patches nearby in the atlas image and computes their similarity to the target image at the given location [95]. These patches are then weighted and the weights can be used as a local weighting for the voxel [96]. This idea was first presented as non-locally weighted vote. Non-local correspondence was then incorporated into the

STAPLE framework as part of non-local STAPLE where non-local correspondence was used as an initial smooth of the labels into the probabilistic STAPLE algorithm [97]. Non-local correspondence is also used in Joint Label Fusion (JLF) [98]. In JLF, non-local correspondence is used to assess the pairwise error rate of two atlases with respect to the truth. Each atlas is then weighted at each location based on a function of the pairwise error rate.

Lastly, atlas selection is an important step in many label fusion algorithms. In atlas selection a set of the available atlases is chosen to use in the algorithm and set of atlases are excluded. Atlas selection can occur both in statistical and weighted voting algorithms. In weighted voting based algorithms, the vote weight for an atlas can be fixed to 0. By setting the atlas's weight globally to zero, this eliminates any influence that atlas might have on any label decision at a given voxel. In statistical fusion algorithms the confusion matrix can be set to a constant value. By setting the confusion matrix to a constant value, the atlas will have no influence to any label estimation during the expectation step of the EM algorithm used to estimate their performance. Then, during the maximization step of EM, the confusion matrix for the atlas is not maximized but instead set to the same constant value.

2.3. Hybrid Segmentation Techniques

Many techniques are available which combine model-based techniques with atlas-based techniques. One example of this is the use of machine learning algorithms to identify common error types in segmentation algorithms [99]. A second example of this is using the probabilistic output of multi-atlas segmentation as the volume to fit an active shape model [100]. Other examples include using graph cuts to correct for segmentation boundaries [101].

2.4. Evaluation of Segmentations

In order to assess the quality of a segmentation algorithm, metrics are needed to assess the accuracy of the algorithm. The first metric is Dice Similarity Coefficient (DSC), which is a measure of overlap between two segmentations [102]. DSC is equal to $\frac{2|T \cap E|}{|T|+|E|}$ where |T| is the number of voxels in the true segmentation, |E| is the number of voxels in the estimated segmentation, and $|T \cap E|$ is the number of voxels where T and E are the same. DSC has a value between 0 and 1 and can be calculated for each label or globally [102]. The second and third metric are mean surface distance and Hausdorff distance [103]. These metrics begin by calculating the surface of the true and estimated segmentation, then they determine a correspondence between the surfaces. Mean surface distance is the average absolute distance between the segmentations. Hausdorff distance is the maximum absolute distance between the segmentations [103]. Mean and Hausdorff distance both have a lower bound of 0, but no upper bound. DSC is criticized because larger structures tend to have higher DSC values since a large portion of the structure can be consensus, for instance the label "white matter" in the human brain suffers from this. MeansSurface and Hausdorff distance do not suffer from this bias.

These metrics can be used in several contexts in medical imaging, for instance comparing two or more segmentation algorithms or for assessing the number of atlases needed for a task. The common method to compare segmentation techniques is to break the atlas population into a training and evaluation set. The training set is used to perform the segmentation and the evaluation set it used to test the accuracy of the segmentation. In the case of evaluating several segmentation algorithms or approaches, the different techniques are calculated on the evaluation set and DSC, mean surface distance, and Hausdorff distance. A Wilcoxon sign-rank test is the calculated

between the results of the different techniques to determine if the results are significantly different. In the case of increasing numbers of atlases, the evaluation set is segmented with increasing numbers of atlases and the atlases are Monte Carlo'd to assess the accuracy with increasing numbers of atlases. This process is inherently biased in that it does not properly model the variance of each atlas and it does not model how each individual evaluation dataset's accuracy is changing.

In conclusion, we have presented algorithms to use jointly multiple labeling protocols for multi-atlas segmentation. We have compared these approaches both in simulation and an empirical study. Our results show statistically significant improvements in comparison to previously published gold-standard techniques when evaluated with defined truth models for the simulation and manually labeled examples for the empirical data.

3. Informatics

At Vanderbilt University Medical Center, the standard practice for clinical care for patients with PD undergoing deep brain stimulation surgery is to receive pre-operative magnetic resonance imaging (MRI) and computed tomography (CT) scanning. The data from these scans are then transferred to the (MIPS) lab via a picture archiving and communication system (PACS) . From there, the data are delivered to the medical and statistical inference (MASI) lab for anonymization and processing.

3.1. MR Imaging and Data Transfer

For each PD patient, the following imaging sequences were scanned. First, a T1-weighted MPRAGE sequence was acquired with TR/TI/TE=7.9/927/3.6ms. Second, a T2-weighted spin echo sequence with TR/TE=3000/80ms. Third, a diffusion weighted sequence was acquired with

32 directions, B-value of 1000, and TR/TE=1000/2.3ms. Fourth, a Fast Gray Matter Acquisition T1 Inversion Recovery (FGATIR) scan with TR/TE/TI=3000/4.39/600ms was acquired. These scans were then transferred using the Digital Imaging and Communications in Medicine (DICOM) standard from Vanderbilt University Medical Center (VUMC) to a PACs .

3.2. Long-term Data Storage and Anonymization

DICOM is inherently an identified format, and can contain protected health information (PHI). For the purposes of research, PHI is not necessary and thus it is important to exclude patient information from research process. The neuroimaging informatics technology initiative (NifTI) file format is a de-identified file format allowing for the storage of medical image volumes. The DICOM data containing PHI is stored on an encrypted hard drive and mounted only for storage and retrieval of files. After storage on the encrypted partition, the DICOM files are converted to NifTI files and BVAL and BVEC files when necessary, using the DICOM toolkit. These files are then uploaded to a project in Vanderbilt's eXtensible Neuroimaging Archive Toolkit (XNAT). XNAT is a database designed with the purpose of storing medical imaging data and facilitating efficient processing on the data.

3.3. Distributed Automation of Image Processing Tasks

Image processing tasks often both have high numbers of parameters and are memory and processor intensive. To preserve consistent results and efficiently complete tasks, systems built to maintain parameter information and distribute tasks across grid and cloud computing environments are needed. At Vanderbilt, the Advance Computing Center for Research and Education (ACCRE) provides an affordable and efficient grid computing environment. In order to interface between

long-term storage, XNAT, and grid computing, ACCRE, the Distributed Automation for XNAT (DAX) was developed to download, store, and process data under these conditions.

DAX operates using two fundamental operators to build, execute, and store data: the processor and the spider. For a given task that we wish to perform, for instance a whole-brain multi-atlas segmentation with the BrainCOLOR protocol, first, a spider is created which performs the task on a particular dataset. Second, a processor is developed which determines whether a particular scan or session is fit to perform a given task. The spider is tasked with the download, processing, and upload of data and the processor is tasked with determining whether a particular dataset has the necessary requirements to execute the task

4. Dissertation Focus

MRI and its applications have made several significant contributions to our understanding of PD. Current MRI imaging and segmentation techniques do not well compensate for changes in imaging sequences, when the atlases do not match the targets and multiple data sets are compared. There is also room for improved segmentation approaches incorporating novel or optimized imaging sequences into atlas-based segmentation approaches. Improved approaches for understanding sub-cortical fiber tracking are needed, since fiber tracts in the sub-cortex are challenging to disentangle in standard imaging.

4.1. Open Problems

Applications of MRI in brain imaging is not a novel concept, but given recent technological and methodological advances, several open problems exist

- Currently, all available segmentation algorithms required all atlases to be segmented with the same protocol, even though there is useful information which can be used between atlases of varying protocols. In PD, there are several structures of interest that there are currently few or no atlases available. Leveraging atlases with different protocols will improve the segmentation results
- Current segmentation approaches inherently assume that the imaging sequence of the target image matches the sequence of the atlas. As the imaging sequence varies, such as the inversion time in an T1-weighted MP-RAGE, the results of the segmentation may change. As a result, retrospective studies comparing PD patients to disease cohorts and healthy controls with different imaging sequences will have an implicit bias form the sequence. Thus, and approach which minimizes the bias between sequences will increase the statistical power of the study.
- Robust and efficient segmentation approaches are not currently available to accurately segment subcortical structures. These subcortical structures are a core focus of research in PD and in order the understand the progression of the disease, better segmentation approaches are needed.
- Segmentation algorithms do not currently support multiple imaging sequences effectively, and thus novel or optimized sequences cannot be incorporated in circumstances when they are useful. Further, segmentation algorithms can also be used to validate and determine which sequences from a given exam card are needed. In PD, many structures have contrast boundaries present in only one modality.

Thus, segmentation approaches incorporating multiple modalities will improve the accuracy.

- Currently, atlas-based segmentation approaches do not properly model variance when evaluating the number of atlases needed for a given segmentation task. In order to build atlases for PD, an understanding of how to properly calculate the number of atlases needed is an important step.
- Large-scale studies of brain development in PD compared with diseases like Alzheimer's disease, essential tremor, and normal aging have currently not been completed. Modern imaging and the availability of large-scale studies is catching up, but techniques are needed to properly correct for these studies.

Here, we propose to address these issues raised by improving on the ideas already available in image segmentation and fiber tracking. In Chapter II, we present a segmentation approach which incorporates multiple labeling protocols into segmentation. In Chapter III, we present an approach for decreasing the effect of having multiple imaging sequences in the study. In Chapter IV, we present an algorithm for efficient segmentation of the hippocampus and amygdala with nearly 200 atlases. In Chapter V, we present an algorithm for segmentation of the cerebellum using atlases which significantly vary in the anatomical presentation. In Chapter VI, we present an algorithm for segmentation of the sub-cortical grey matter using multiple imaging sequences. In Chapter VII, we present an algorithm for proper estimation of variance in multi-atlas segmentation with respect to the number of atlases used.

4.2. Contributions

- We designed a segmentation approach which incorporates multiple labeling protocols into segmentation. This approaches uses a generalization of the confusion matrix in STAPLE to incorporate differing labeling protocols.
- We presented an approach for decreasing the effect of having multiple imaging sequences in the study. This approach synthesizes atlases with the target imaging sequence utilizing atlases with underlying biological parameter maps.
- We present an algorithm for efficient segmentation of the hippocampus and amygdala with nearly 200 atlases. This approach uses a reduced field of view segmentation where the registration and segmentation for the multi-atlas segmentation are completed only on the region surrounding the hippocampus and amygdala.
- We present an algorithm for segmentation of the cerebellum using a diverse population of atlases and strong atlas selection. This approach provides each atlas patch with its own confusion matrix, so that rater performance is spatially evaluated.
- We present a segmentation algorithm for the evaluation and applicability of multiple imaging sequences in the sub-cortex. This approaches uses multi-modal registration and segmentation and breaking the segmentation problem into individual problems for each region of interest.

- We present an algorithm for proper estimation of variance in multi-atlas segmentation with respect to the number of atlases used. This approach uses monte-carlo calculation of the underlying distribution of variance.

5. Narrative

This work centers on applications of image processing to the study of PD. At the beginning of this dissertation, multi-atlas segmentation was becoming a well-characterized approach when working on a restricted atlas population. In order to translate these approaches to PD, several important steps needed to be taken. The imaging sequences present in PD cases are not identical to the datasets available in other cases. Thus, a segmentation approach which was aware of imaging sequence of the target image was developed. Also, a segmentation algorithm for multi-modal segmentation of particular structures of interest was developed to incorporate the specific sequences commonly available in PD.

There are several structures of interest in PD which are not of interest in other diseases or conditions. Thus, specialized atlases are needed to segment these structures. Two approaches were considered to improve our understanding of how to build atlases quickly. First, an atlas re-use approach was developed where atlases with a variety of labeling protocols could be used in a multi-atlas segmentation framework. This is of interest in PD because it may decrease the number of newly labeled atlases needed for a given task. Second, an approach to properly characterize variance with respect to the number of atlases used in a segmentation task was considered. This is also important because it helps to better determine the number of atlases needed for a segmentation approach.

In order to characterize the structures of interest in PD, several specialized segmentation approaches were developed. All of these segmentation approaches used localized segmentation of the structure of interest. Particularly, segmentation approaches for the hippocampus and amygdala, cerebellum, and subcortical brain structures, were developed. Each approach utilizes different optimization strategies to address particular concerns within that structure of interest.

Chapter II

Multi-Protocol, Multi-Atlas Statistical Fusion: Theory and Application

1. Introduction

The multi-atlas technique has become an essential medical image processing approach and been adopted widely for applications ranging from the brain [84] to the abdomen [104] and pelvic structures [105]. The promise of generalizing robust algorithms from limited collections of labeled data without needing to posit specific structural models is highly appealing for clinical applications and rapid prototyping. However, manual labeling of medical images can be extraordinarily resource intensive. For each new application (or even refinement of an existing application), multi-atlas methods require labeling a new atlas set.

Consider segmentation of the hippocampus. Numerous protocols (e.g., [106]) exist to delineate the hippocampus in MR images. These protocols vary on the basis of hippocampal white matter, the border between the hippocampus and amygdala, the hippocampal tail, and various other markers. With current techniques, individuals interested in studying the hippocampus are limited to either ignoring all other label sets or using a coarse protocol as an anatomical "stamp" and are left to rectify the space between the protocols. This is clearly suboptimal as it does not allow for joint inference between protocols and does not specifically estimate any one protocol.

Recently, [107] illustrated that protocols need not be considered fully independent and that a generative latent model could be used to exploit the dependence between protocols. However, [107] required a specific human-provided mapping function to join protocols. Herein, we revisit

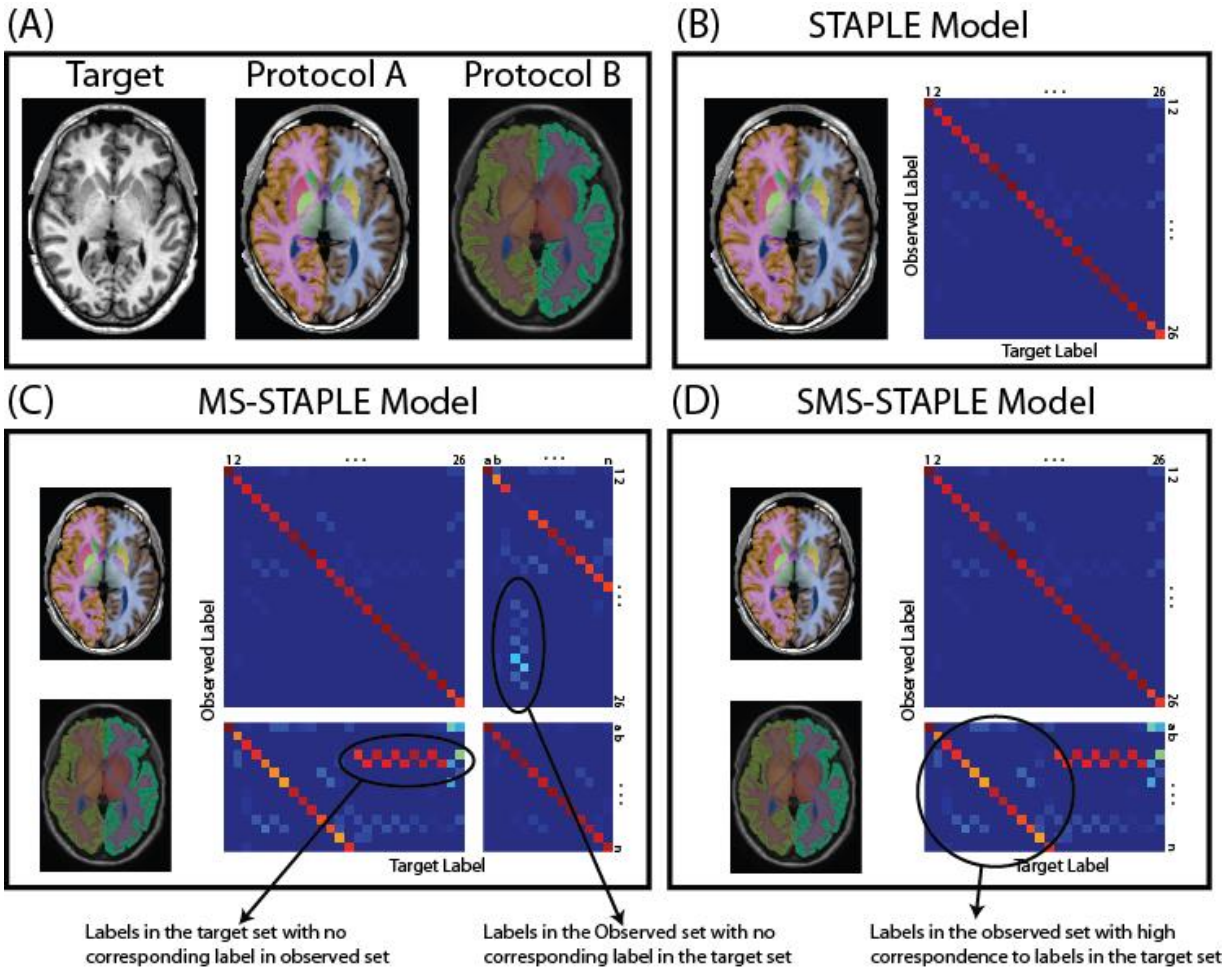

Figure II-1 Modeling rater performance with the STAPLE, MS-STAPLE, and SMS-STAPLE algorithms. A target (A) image can be labeled using atlases of multiple protocols. In STAPLE (B) only atlases with a 1:1 correspondence can be used in segmentation, in MS-STAPLE (C) atlases of multiple protocols can be used to jointly segment *all* of the protocols, and in SMS-STAPLE (D) atlases of multiple protocols can be used to jointly segment *one* target protocol.

the need to combine atlases of different protocols and show that dependence structure between atlases can be learned without human intervention. Our approach leads natural generalization of majority vote and locally weighted vote (while maintaining computational efficiency). The dependence structure can also be estimated within the Simultaneous Truth and Level Estimation (STAPLE) perspective through jointly modeling the protocol behaviors.

This manuscript is organized as follows. Section 2 presents notation and derives the theory for fusion. Section 3 develops and characterized the performance of the methods in a simulated example with a well-known error model and in an empirical example of deep brain labeling. Section 4 concludes with a brief discussion.

2. Theory

We first present the underlying framework connecting generalization of voting fusion approaches and MS-STAPLE. Consider a target image with N voxels and intensity values $\boldsymbol{I} \in \mathbb{R}^{N\text{x}1}$ with a set of R registered atlases, associated intensities $\boldsymbol{A} \in \mathbb{R}^{N\text{x}R}$, and registered labels $\boldsymbol{D} \in \boldsymbol{\mathcal{L}}^{N\text{x}R}$. Let D_{ij} be the decision of atlas j at voxel i. In the case of multiple sets of labeling protocols, $\mathcal{L}$ corresponds to an arbitrary set of labels, that is to say that the numerical label denotations between the atlases do not necessarily correspond to the same anatomical structure. Lastly, let $\boldsymbol{L} = \{L_1, L_2, \dots, L_\psi\}$, where L_j corresponds to the number of labels in atlas class j and ψ corresponds to the total number of different labeling protocols used with a vector $\boldsymbol{d} \in \mathbf{I}^{R\text{x}1}$ mapping each rater to its atlas class. The objective of label fusion is to estimate $\hat{S}$, a voxel discrete mapping of the target image. Figure 1 illustrates the rater models from the following sections.

2.1. Generalized Majority, Locally Weighted, and Non-Locally Weighted Vote

In standard majority vote (MV), the probability of label s at voxel i, $p_{MV,i}(s|D)$, is empirically determined as the fraction of atlases that observe s at voxel i. We may generalize MV to include atlases (raters) of different protocols by:

$$p_{GMV,i}(s|D) = \frac{1}{R}\sum_j p_i(s|D_{ij}, d_j) \tag{1}$$

where $p_i(s|D_{ij}, d_j)$ is discrete probability of the co-occurrence across labeling protocols of registered atlases (which can be empirically estimated as seen in §2.4). Henceforth, for convenience, we simplify the co-occurrence probability to be spatially invariant and drop the subscript. Generalized majority vote (GMV) is thus $\widehat{S_{GMV,i}} = \underset{s}{\operatorname{argmax}}\, p_{GMV,i}(s|D)$. Similarly, locally weighted vote [84] can be framed around the co-occurrence probability as:

$$p_{LWV,i}(s|D, d, A, I) = \frac{1}{Z}\sum_j p_i\left(s\middle|D_{ij}, d_j, A_{ij}, I_i\right) = \frac{1}{Z}\sum_j p\left(A_{ij}\middle|I_i\right)p\left(s\middle|D_{ij}, d_j\right) \quad (2)$$

assuming marginal independence of image intensity values and observed labels, where Z is a partition function normalizing distribution to a valid probability distribution and $p\left(A_{ij}\middle|I_i\right)$ is the likelihood of observing the intensity value of atlas j at voxel i. Hence, $\widehat{S_{LWV,i}} = \underset{s}{\operatorname{argmax}}\, p_{LWV,i}(s|D_i, d, A, I)$.

Moreover, non-locally weighted voting (NLWV) [96, 97] can be reframed with $p_{NLWV,i}(s|D, d, A, I) = \frac{1}{N}\sum_j \sum_{s' \in L_{d_j}} p_{i,j}(s'|D, d, A, I)p(s|s', d_j)$ with s' as the latent correspondence and $p_{i,j}(s'|D, d, A, I)$ as the likelihood of atlas j observing s' at i. Note that the latent corresponding label, s', occupies the role of the atlas labels in the NLWV model, and $\widehat{S_{NLWV,i}} = \underset{s}{\operatorname{argmax}}\, p_{NLWV,i}(s|D_i, d, A, I)$.

Note that these formulations reduce to their classical definitions if all of the atlases are of the target class (i.e., co-occurrence matrix are 1:1) or the co-occurrence probabilities of non-target classes are uniform (i.e., other label protocols are uninformative and, hence, ignored).

2.2. Multi-Set STAPLE (MS-STAPLE)

STAPLE label fusion maintains a confusion matrix θ for each atlas [91]. Each confusion matrix entry $\theta_{js's} \equiv p(D_{ij} = s' | T_i = s)$ presents a discrete probability distribution where s' corresponds to the label observed and T corresponds to a latent true segmentation. Consider segmentation with ψ labeling protocols, each with its own confusion matrix such that the likelihood of observing a label is $f(D_{ij} = s' | T_i = l, \{\theta_j\})$ where s' is the observed later by rater j at voxel i, l is a set of true labels of size 1xψ observed at i, and $\{\theta_j\}$ is a set of ψ confusion matrices for atlas j where $\theta_{j\rho s's}$ corresponds to the likelihood that label s' observed by rater j given that label s of set ρ is the true label. Note that $\theta_{j\rho}$ is a possibly non-square matrix where $\theta_{j\rho}$ is of size L_{d_j}xL_ρ. This is the core of MS-STAPLE.

To estimate the data likelihood, we follow [93] to capture dependence between protocols through a geometric mean:

$$f(D_{ij} = s' | T_i = l, \{\theta_j\}) = \left(\prod_{\rho \epsilon \psi} f(D_{ij} = s' | T_{i\rho} = l_\rho, \theta_{j\rho}) \right)^{\alpha_{jl}}$$

$$= \left(\prod_{\rho \epsilon \psi} \theta_{j\rho s' l_\rho} \right)^{\alpha_{jl}} \quad (3)$$

where a_{jl} maintains $\sum_{s' \in L_{d_j}} \left(\prod_{\rho \epsilon \psi} \theta_{j\rho s' l_\rho} \right)^{a_{jl}} = 1$, thus maintaining a valid probability distribution. Note that [93] captured joint information across specific hierarchical protocols, while here we are using it to normalize across relationships that must be estimated assuming conditional independence between sets of labels.

Given this model, we can apply expectation maximization (EM). Briefly, let $W \in \mathbb{R}^{\prod_{\rho\epsilon\psi} L_\rho \mathrm{x} N}$ where $W_{li}^{(k)} \equiv f\left(T_i = l \middle| D, \{\theta\}^{(k)}\right)$ is the probability that the true label set observed at voxel i during iteration k is l. Using a Bayesian expansion and the assumed conditional independence between atlases,

$$W_{li}^{(k)} = \frac{f(T_i = l) \prod_{j\epsilon R} f\left(D_{ij} = s' \middle| T_i = l, \left\{\theta_j^{(k)}\right\}\right)}{\sum_{l'} f(T_i = l') \prod_{j\epsilon R} f\left(D_{ij} = s' \middle| T_i = l', \left\{\theta_j^{(k)}\right\}\right)} \tag{4}$$

where $f(T_i = l)$ is a voxelwise *a priori* distribution of the underlying segmentation. The denominator corresponds to a partition function normalizing $\boldsymbol{W}$ to a valid probability distribution. Substituting (3) in (4) yields the MS-STAPLE E-Step,

$$W_{li}^{(k)} = \frac{f(T_i = l) \prod_{j\epsilon R} \left(\prod_{\rho\epsilon\psi} \theta_{j\rho s' l_\rho}^{(k)}\right)^{a_{jl}}}{\sum_{l'} f(T_i = l') \prod_{j\epsilon R} \left(\prod_{\rho\epsilon\psi} \theta_{j\rho s' l'_\rho}^{(k)}\right)^{a_{jl'}}} \tag{5}$$

To estimate the performance parameters, maximize the expected value of the conditional log-likelihood function. We follow the traditional M-Step expansion:

$$\begin{aligned}
\theta_{j\rho}^{(k+1)} &= \underset{\theta_{j\rho}}{\operatorname{argmax}} \sum_i E\left[\ln\left(f\left(D_{ij} \middle| Ti, \{\theta_j\}\right)\right) \middle| D, \{\theta_j\}\right] \\
&= \underset{\theta_{j\rho}}{\operatorname{argmax}} \sum_{s'} \sum_{i:D_{ij}=s'} \sum_l W_{li}^{(k)} \ln\left(\prod_{\rho\epsilon\psi} \theta_{j\rho s' l_\rho}^{(k)}\right)^{\alpha_{jl}^{(k)}} \\
&= \underset{\theta_{j\rho}}{\operatorname{argmax}} \sum_{s'} \sum_{i:D_{ij}=s'} \sum_l W_{li}^{(k)}\, \alpha_{\mathrm{jl}}^{(\mathrm{k})} \sum_{\rho\epsilon\psi} \ln\left(\theta_{j\rho s' l_\rho}^{(k)}\right)
\end{aligned} \tag{6}$$

where $i: D_{ij} = s'$ corresponds to the voxels where $D_{ij} = s'$. To constrain each row of the confusion matrix to be a valid probability distribution ($\sum_{s'} \theta_{j\rho s's} = 1$), we differentiate by each element and use a Lagrange Multiplier (λ):

$$0 = \frac{\partial}{\partial \theta_{j\rho s's}} \left[\sum_{s'} \sum_{i:D_{ij}=s'} \sum_{l} W_{li}^{(k)} \alpha_{lj}^{(k)} \sum_{\rho \epsilon \psi} \ln\left(\theta_{j\rho s' l_\rho}^{(k)}\right) + \lambda \sum_{s'} \theta_{j\rho s's}^{(k)} \right]$$

$$0 = \sum_{i:D_{ij}=s'} \sum_{l:l_\rho=s} \left(\frac{W_{li}^{(k)} a_{jl}^{(k)}}{\theta_{j\rho s's}} \right) + \lambda$$

$$-\lambda \theta_{j\rho s's}^{(k+1)} = \left(\sum_{i:D_{ij}=s'} \sum_{l:l_\rho=s} W_{li}^{(k)} a_{jl}^{(k)} \right)$$

$$\theta^{(k+1)}{}_{j\rho s's} = \frac{\left(\sum_{i:D_{ij}=s'} \sum_{l:l_\rho=s} W_{li}^{(k)} a_{jl}^{(k)} \right)}{\left(\sum_{i} \sum_{l:l_\rho=s} W_{li}^{(k)} a_{jl}^{(k)} \right)} \tag{7}$$

where $l: l_\rho = s$ corresponds to label sets where the voxel of atlas set ρ is s.

2.3. Simplified MS-STAPLE (SMS-STAPLE) and Non-Local-SMS-STAPLE

Empirically, we have found the fully parameterized MS-STAPLE model (§2.2) less numerically stable than one would desire. Here, we present a simplified model to improve stability with limited data. In place of $\{\theta_j\}$ (with $\sum_j L_{d_j} \Pi_\rho L_\rho$ degrees of freedom), consider $\tilde{\theta}_j$, a $L_{d_j} \times L_t$ confusion matrix with t to be the index of the target label. Each element of $\tilde{\theta}_{js's}$ corresponds to the probability rater j observes label $s' \in L_{d_j}$ given that the true label is $s \in L_t$. Note that each atlas has confusion matrix with the number of rows dependent on the labeling protocol, but the

number columns matching the target protocol. The degree of freedom of $\tilde{\theta}$ is $\sum_j L_{d_j} L_t$, a possibly dramatic reduction from the §2.2 model by assuming conditional independence.

With Simplified Multi-Set STAPLE model (i.e., SMS-STAPLE), the EM update equations are found to be:

$$W_{si}^{(k)} = \frac{f(T_i = s) \prod_j \tilde{\theta}_{js's}^{(k)}}{\sum_{s''} f(T_i = s'') \prod_j \tilde{\theta}_{js's''}^{(k)}} \tag{8}$$

for the E-Step, and

$$\tilde{\theta}_{js's}^{(k+1)} = \frac{\left(\sum_{i:D_{ij}=s'} W_{si}^{(k)}\right)}{\left(\sum_i W_{si}^{(k)}\right)} \tag{9}$$

for the M-Step.

Following [97], we derive the EM update equations to incorporate non-local correspondence in the SMS-STAPLE model (herein SMS-Non-Local STAPLE) as:

$$W_{si}^{(k)} = \frac{f(T_i = s) \prod_j \sum_{i' \in \aleph_S(i)} \tilde{\theta}_{js's}^{(k)} c_{ji'i}}{\sum_{s''} f(T_i = s'') \prod_j \sum_{i' \in \aleph_S(i)} \tilde{\theta}_{js's''}^{(k)} c_{ji'i}} \tag{10}$$

for the E-Step where $\aleph_S(i)$ is the spatial neighborhood around voxel i and $c_{ji'i}$ is the likelihood of correspondence between atlas j at voxel i' and the target at voxel i. The M-Step follows as:

$$\tilde{\theta}_{js's}^{(k+1)} = \frac{\sum_i \left(\sum_{i' \in \aleph_S(i):D_{i'j=s'}} c_{ji'i}\right) W_{si}^{(k)}}{\sum_i W_{si}^{(k)}} \tag{11}$$

2.4. Modeling Co-Occurrence Probability and Initialization

For the voting approaches (§2.1), the co-occurrence probability definitions were calculated empirically as

$$p(s|s',\rho) = \frac{\sum_{j:d_j=\rho}\sum_{h:d_h=t}\sum_{i:D_{ij}=s'}\delta(D_{ih},s)}{\sum_{j:d_j=\rho}\sum_{i}\delta\left(D_{ij},s'\right)} \tag{12}$$

where s is the latent label, s' is the observed label, and ρ is the index of the label set from which s' is drawn. Note that this model is conditionally independent of the true label. For the MS-STAPLE approaches (§2.2) were initialized by:

$$\theta^{(0)}{}_{j\rho s's} = \theta^{(0)}_{d_j\rho s's} = \frac{\sum_{k:d_k=\rho}\sum_{h:d_h=t}\sum_{i:D_{ik}=s'}\delta(D_{ih},s)}{\sum_{k:d_k=t}\sum_{i}\delta(D_{ik},s)} \tag{13}$$

For the SMS-STAPLE approaches (§2.3), initial confusion matrixes were:

$$\theta^{(0)}_{js's} = \theta^{(0)}_{d_js's} = \frac{\sum_{j:d_j=\rho}\sum_{h:d_h=t}\sum_{i:D_{ij}=s'}\delta(D_{ih},s)}{\sum_{j:d_j=t}\sum_{i}\delta\left(D_{ij},s\right)} \tag{14}$$

Convergence of EM was detected when the average change in $\tilde{\theta}$ from k to $k+1$ was less than $\epsilon = 10^{-6}$, specifically:

$$\frac{1}{M}\sum_{j}\sum_{s'\in L_{d_j}}\sum_{s\in L_t} abs\left(\tilde{\theta}^{(k+1)}_{js's} - \tilde{\theta}^{(k)}_{js's}\right) \tag{155}$$

where M is the total number of elements in all of the confusion matrices (i.e. $\sum_j L_{d_j} \mathrm{x} L_t$).

3. Methods and Results

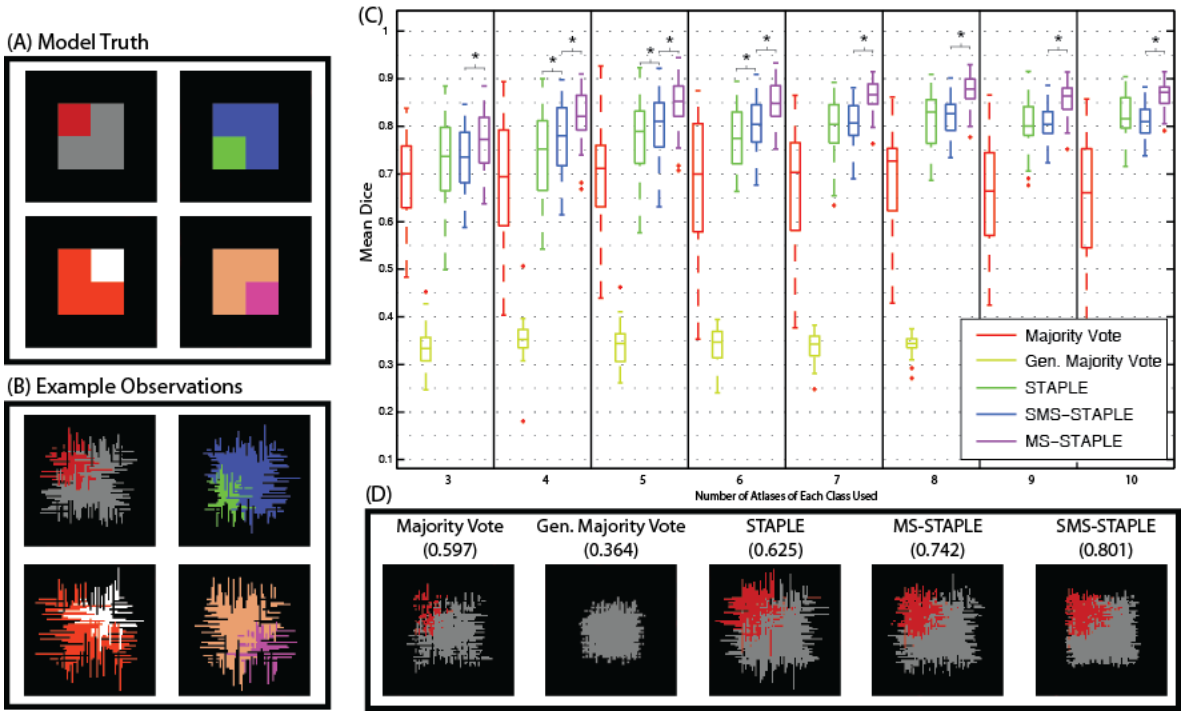

Figure II-2 Simulation results based on algorithm and number of atlases used. A set of simulated truth models (A) were generated to model a relationship with inter-protocol spatial dependence. Simulated observations with increasing numbers of observations (B) were fused (C and D). For STAPLE v. MS-STAPLE and SMS- v. MS-STAPLE: * $p < 10^{-5}$.

3.1. Simulation of Distinct Protocol

We first consider labeling of an idealized square object within an 80x80 pixel background where the object consists of four distinct quadrants. There were four possible labeling protocols, and each protocol labels one quadrant as distinct from the other three (see Figure 2A). The simulated raters independently made errors in their respective protocols in terms of randomly shifted boundaries, which has been commonly used since [92]. Boundary shift errors were selected uniformly at random from between -10 and +10 pixels for boundary point (Figure 2B). Here, we evaluated fusion of between three and ten raters for each protocol (i.e., between 12 and 40 total

simulated rater label sets) for a single target image. The experiment was Monte Carlo repeated with new simulated rater label sets 20 times to establish model variability.

Each of the fusion algorithms (§2.1-2.3) was applied to each unique set of simulated rater observations (Figure 2C). Note that the distinct algorithms were applied to same simulated data for each number of raters and Monte Carlo iteration. Performance was evaluated by Dice similarity coefficient and Wilcoxon sign rank test. MS-STAPLE showed significant improvement over STAPLE with four to six training examples of each class available ($p<10^{-5}$). SMS-STAPLE showed significant improvement over all models with all numbers of simulated observations ($p<10^{-5}$). Generalized Majority Vote was significantly worse than all models ($p<10^{-5}$).

3.2. Empirical Deep Brain Segmentation

To evaluate empirical performance, we constructed a dataset with two distinct whole-brain protocols (a fine protocol and a coarse protocol). To ensure that true results were well known, we studied 40 T1-weighted MRIs expertly labeled with 14 fine deep brain structures and 12 coarse labels for the remainder of the brain (derived from the BrainCOLOR protocol; Neuromorphometrics, Inc., Somerville, MA). For ten randomly selected subjects, we reduced the deep brain structures to two lateralized labels (i.e., the coarse protocol), and the remaining 30 subjects were randomly split into two groups of 15, one for training and one for testing (both with the fine protocol). All pairs of images were co-registered with ANTS-Syn with default parameters [88], and labels were deformed to match the target images with nearest neighbor interpolation.

We evaluated a situation where the ten coarse atlases we assumed to be preexisting and between two and 15 of the fine training atlases were made available. For each number of new fine atlases (Q), we randomly selected Q atlases of the 15 training set to construct a simulated available

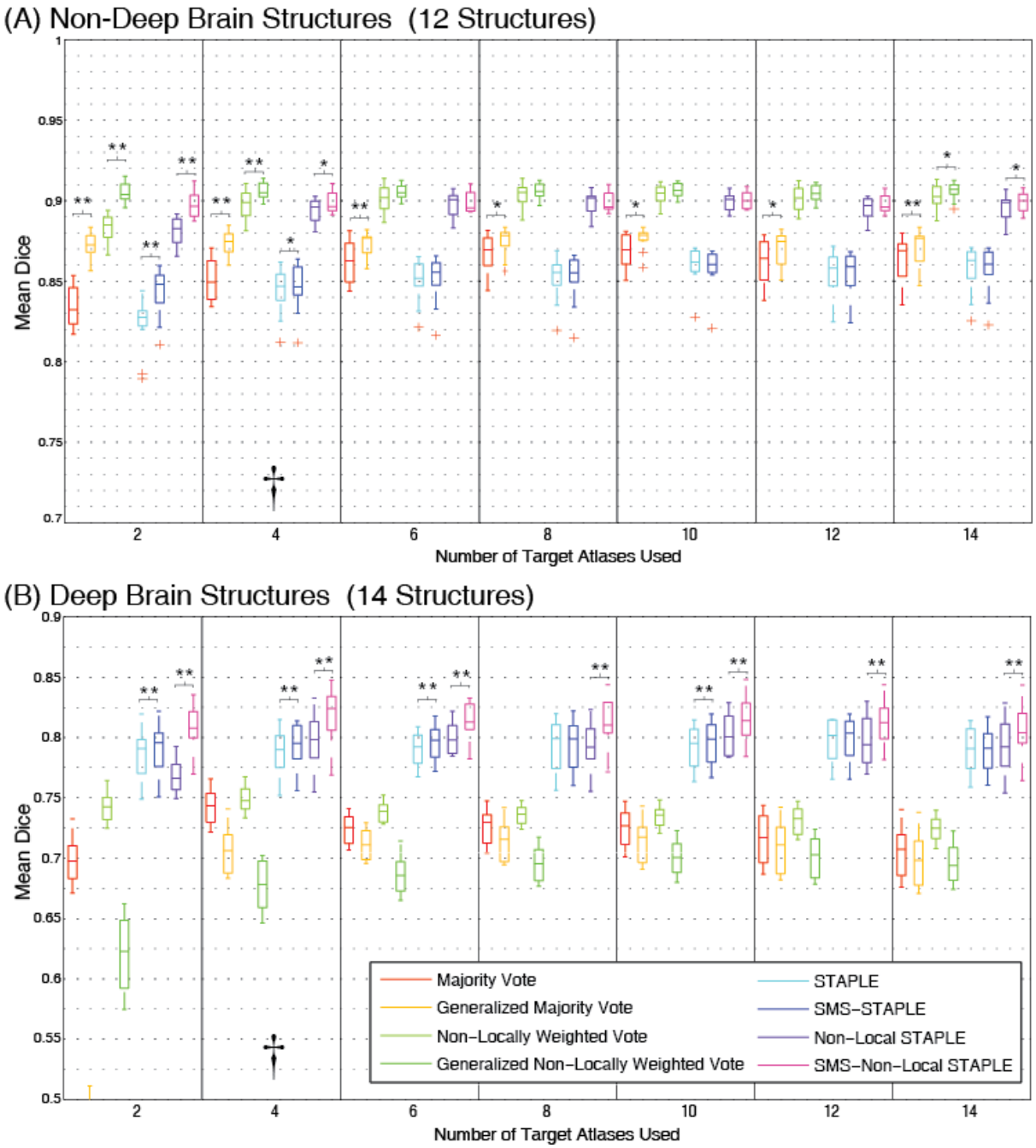

Figure II-3 Results from empirical deep brain segmentation experiment. Both SMS-STAPLE and SMS-Non-Local STAPLE show statistically significant improvement (*$p < 0.05$, ** $p < 0.01$) for many of the segmentation tasks particularly when few atlases of the target class are available. † indicates segmentation task shown in Figure 4.

dataset of 10 coarse and Q fine atlases. For each of the 15 images in the testing set, we performed

each of label fusion algorithms and evaluated performance with the Dice similarity coefficient and statistically evaluated with the Wilcoxon sign-ran k test (Figure I-3).

For non-deep brain structures (Figure 3A), generalized majority vote outperforms majority vote with all numbers of atlases, and generalized non-locally weighted vote, SMS-STAPLE, and SMS-Non-Local STAPLE outperform their counterparts when few (<6) target atlases are available ($p<0.05$). For deep-brain structures (Figure 3B), the majority vote and local weighted vote outperformed generalized majority vote and generalized locally weighted vote ($p<0.05$). Meanwhile, SMS-STAPLE outperformed STAPLE when few target atlases were available and SMS-Non-Local STAPLE outperformed Non-Local STAPLE in all experiments ($p<0.01$). Note that MS-STAPLE results are not shown and performance was worse than all methods (as discussed in the theory). These results are shown qualitatively in Figure 4.

4. Discussion and Conclusion

In manuscript presents generalizations of majority vote and locally weighted vote to incorporate multiple labeling protocols into the segmentation of a target label set. We also present a generalization of the STAPLE algorithm to jointly incorporate and segment multiple protocols and a simplification of this model for an individual target protocol. To achieve tractable models, we assume conditional independence between the labeling protocols. As an aside, an alternative approach would have been to design a "wide" confusion matrix θ_j of size $L_{d_j} \mathrm{x} \prod_{\rho \in \psi} L_\rho$ per rater to capture all potential relationships, but the degrees of freedom in such approach quickly exceed the number atlases likely to be available in practice and the model would reduce to classic STAPLE separately for each protocol unless atlases labeled with multiple protocols were included.

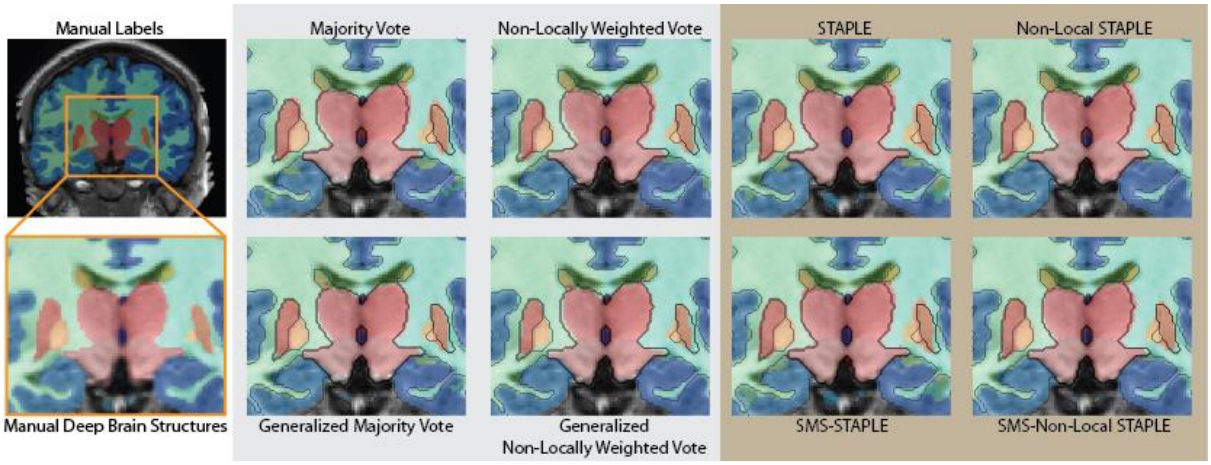

Figure II-4 Qualitative results comparing eight segmentation algorithms (identified by † in Figure 3). Contour lines indicate manual truth target labels (shown lower left).

In [107], the authors develop an alternate formulation of voting based fusion and STAPLE for multiple protocols by defining a matching between the observed labels of the individual protocols, an intermediate protocol specific coarse labeling, and a set of fine labels which they wish to segment. To achieve the fusion, the authors build a mapping of the vector relationship between an individual protocol and the target, assuming there is a clear relationship that can be defined between the protocols, which in many cases may not be achievable.

In comparison, the work presented here assumes a softer relationship between labels. In our empirical study, we found that 46.7% of pairwise label relationships contained non-zero co-occurrence probabilities. In contrast, only 6.4% of the joint label relationships would be captured by [107]. By using pairwise label relationships we are both simplifying the initialization of the model and potentially capturing more robust relationships found within the data.

Lastly, the work of [107] presents a natural generalization of STAPLE, which directly follows directly from the approaches of [91]. The generalizations of [107] assumes a hierarchically defined relationship between the observed labels of each atlas and the target labels. Conversely the

work presented in this dissertation provides an alternate generalization where we jointly segment multiple protocols and learn the relation between them as a function of the observations. We then simplify the proposed model to a more tractable solution for the scale of training examples available.

This work has show how atlases with complementary labeling protocols can be used to improve segmentation of deep brain structures. These techniques hold substantial promise for improving fusion with existing protocols by using related atlases sets from disparate research groups (e.g., from the extensive list of available atlases of distinct protocols, http://www.mindboggle.info/data.html). Alternatively, this approach could form the basis of a bootstrapping technique where a new protocol is developed by extending/refining one or more existing protocols, as is illustrated in the empirical deep brain explain. Finally, the technique could be used to work towards consensus protocols while quantifying the joint information between the protocols. For example, hippocampal sub-field segmentation [106] is a active area of research, but groups providing high resolution protocols are rarely concerned with detailed manual whole-brain labeling. This approach could be used to jointly segment full brain [108] and hippocampus subfield labels without having one set of humans label both.

Chapter III

Synthetic Atlases Improve Segmentation Consistency between T1-Weighted Imaging Sequences

1. Introduction

The intensities of T1-weighted magnetic resonance images are non-quantitative in that they are related to the local T1 relaxation, T2 relaxation, and proton density properties via imaging sequence parameters, but not directly specific individual tissue characteristics [73]. Multi-atlas segmentation (MAS) [109] is commonly used to quantify T1-weighted MRI through voxel-wise segmentation, which can be used to perform volumetric analysis [110], fMRI correlations [111], and tractography [112]. The relationship between intensity and T1-weighted imaging is defined by the parameters of the imaging sequence, e.g., for magnetization prepared rapid gradient echo (MPRAGE) - the inversion time (TI), repetition time (TR), echo time (TE), and flip angle (α).

Imaging sequence plays a significant role in quantitative segmentation results [113, 114], as illustrated in Figure 1. Volume variability is minimized by defining a selective range of protocols with which all subjects are scanned [115]. Within a single site study, researchers typically define a single protocol for all subjects. In larger, multi-site studies, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI), a range of sequence parameters is defined to ensure data consistency between sites [115]. Other recent studies, e.g., the Autism Brain

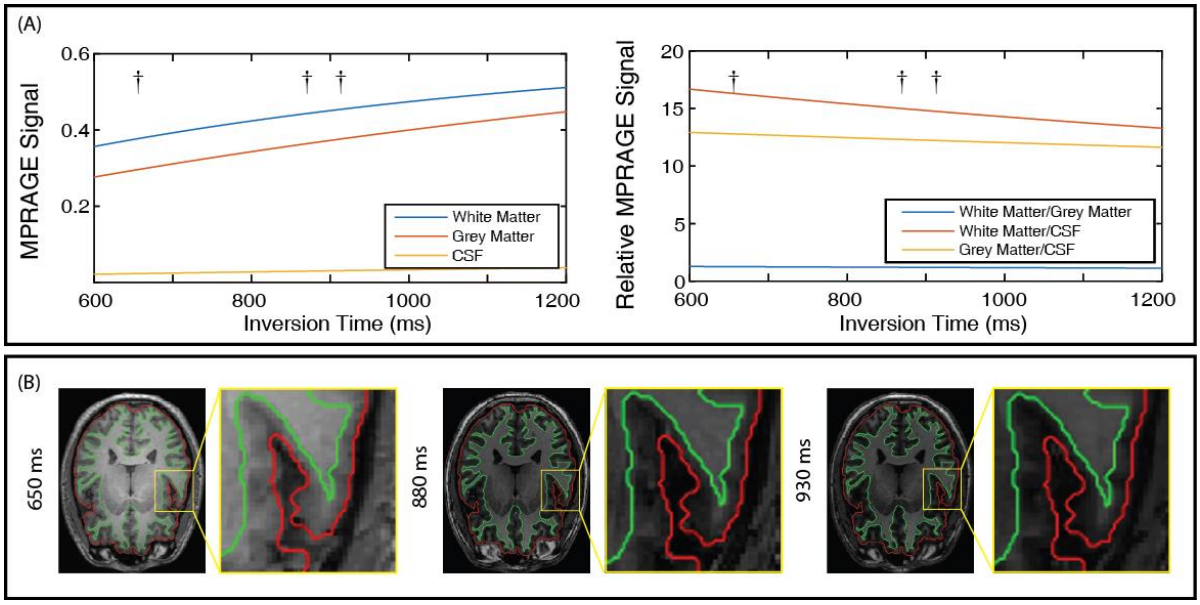

Figure III-1 (A) Absolute and relative MPRAGE signals vary non-linearly based on the sequence parameters. (B) These changes introduce visually apparent boundary shifts when the same structures are compared across different imaging sequences.

Imaging Data Exchange (ABIDE), employ a unique sequence at each site in the study [116]. Similarly, retrospective studies do not have direct control over the sequences used [117]. Large studies (e.g., [117, 118]) may lose statistical power as a result of the added variance in segmentation results or suffer from bias problems if the aspects of the subject population are associated with imaging sequence.

Image synthesis has shown promise results for harmonizing T1-weighted imaging across sequences [119, 120]. To date, the underlying assumption for image synthesis is that one has paired template images. Specifically, for a given target sequence, one has a template with image pairs with both the source and target sequences on at least one individual, albeit on a different individual than the synthesis target. Hence, image synthesis enables translation between two or more sequences that have been simultaneously acquired.

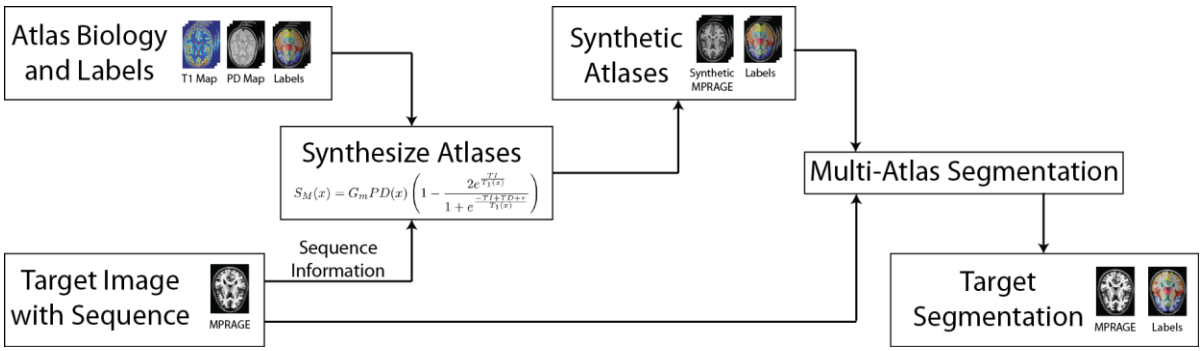

Figure III-2 Atlases with quantitative T1-relaxation and proton density maps are used to generate synthetic atlases from sequence parameters (i.e., inversion time, echo time, repetition time). The synthesized atlases are used in the standard multi-atlas segmentation procedure.

Herein, we focus on the situation in which we have multiple different modalities, but we are not able to acquire paired templates for all modalities. This would be the situation for some multi-site studies and for retrospective studies. As a case study, we focus on the ABIDE study which has 17 distinct imaging sequences. A similar problem has been addressed in the image synthesis literature where the authors estimate the sequence parameters of a given scan and then transform their atlas using imaging physics equations before synthesis to better match the target [119]. The authors produce synthetic target images more similar to their atlas, thus decreasing the error due to sequence variation. This approach is interesting but not directly applicable to multi-atlas segmentation since we cannot assume we have atlas image pairs similar to the target image of multi-atlas segmentation. In this work we consider a similar underlying principle where we first use the imaging physics equations to synthesize atlases more similar to the target and use the new "synthetic" atlases for multi-atlas segmentation.

2. Theory

In this section, we first derive atlases with the underlying physiologic data necessary to synthesize atlases, with a focus on MPRAGE T1-weighted MRI data. Then, we propose a synthetic

multi-atlas segmentation pipeline using the physiologic parameter maps to segment a target scan for which we know the underlying imaging sequence (Figure 2).

2.1. Synthetic Atlas Theory

Signal intensity in MPRAGE images is approximately [73, 119]:

$$S_M(x) = G_M PD(x)\left(1 - \frac{2e^{-\frac{TI}{T_1(x)}}}{1 + e^{-\frac{TI+TD+\tau}{T_1(x)}}}\right) \quad (1)$$

where $S_M(x)$ is the signal intensity at a voxel x, G_M is the scanner gain field, $PD(x)$ is the proton density at x, TI is the inversion time, $T_1(x)$ is the T1-relaxation at x, T_D is the sequence delay time, and τ is the slice timing. The gain field is assumed to be a global scalar, which corresponds to performing spatial bias correction prior to analysis. Volumetric segmentation and registration procedures typically normalize image intensity in the correspondence calculation [97, 121, 122], thus it is not important to explicitly calculate the gain field.

To derive the proton density and T1-relaxation maps, we use a T1-weighted MPRAGE along with a T2-weighted dual echo. The dual echo signal equation is [123]:

$$S_{DSE_n}(x) = G_{DSE} PD(x)\left(1 - 2e^{-\frac{TR-\frac{TE_1+TE_2}{2}}{T_1(x)}} + 2e^{-\frac{TR-\frac{TE_2}{2}}{T_1(x)}} - e^{-\frac{TR}{T_1(x)}}\right)e^{-\frac{TE_n}{T_2(x)}} \quad (2)$$

where n corresponds to the nth echo, G_{DSE} is the dual echo gain field, TR is the repetition time, TE_n is the nth echo time, and $T_2(x)$ is the T2-relaxation at x. Since the repetition time of these scans is long, the T1-weighted component equates to effectively 0 and the equation reduces to

$$S_{DSE_n}(x) \approx G_{DSE} PD(x) e^{-\frac{TE_n}{T_2(x)}} \quad (3)$$

Thus, the local T2 relaxation is

$$T_2(x) \approx \frac{TE_1 - TE_2}{\ln\left(S_{DSE_2}(x)\right) - \ln\left(S_{DSE_1}(x)\right)} \tag{4}$$

which is a standard logarithmic T2-relaxation fit. The local proton density is then found as

$$G_{DSE}PD(x) = \frac{S_{DSE_1}}{e^{-\frac{TE_1}{T_2(x)}}} \tag{5}$$

which is within a scalar factor of the proton density. From this result, the T1-relaxation map can be solved from (1). First, the gain field is solved by

$$\frac{S_M(x)}{G_{DSE}PD(x)} = \frac{G_M PD(x)}{G_{DSE}PD(x)}\left(1 - \frac{2e^{-\frac{TI}{T_1(x)}}}{1 + e^{-\frac{TI+TD+\tau}{T_1(x)}}}\right) \tag{6}$$

$$S_N(x) = G_N\left(1 - \frac{2e^{-\frac{TI}{T_1(x)}}}{1 + e^{-\frac{TI+TD+\tau}{T_1(x)}}}\right) \tag{7}$$

where $S_N(x)$ is a simplification of $\frac{S_M(x)}{G_{DSE}PD(x)}$ and G_N is $\frac{G_M}{G_{DSE}}$. $G_N$is

$$G_N = \frac{S_N(x)\left(1 + e^{-\frac{TI+TD+\tau}{T_1(x)}}\right)}{\left(1 - 2e^{-\frac{TI}{T_1(x)}}\right)} \tag{8}$$

Though this equation still contains and unknown, $T_1(x)$, G_N can be solved with known values of $T_1(x)$. To do this, the value of G_N is solved with the expected values of $T_1(x)$ in the grey matter, white matter, and CSF and the median intensity of these tissue types [124]. The final value of G_N is calculated as the mean of these three measurements. Lastly, to estimate $T_1(x)$ it is necessary to assume that $e^{-\frac{TI+TD+\tau}{T_1(x)}} \approx e^{-\frac{TI}{T_1(x)}}$, which a reasonable assumption since the inversion time is on the order of hundreds of milliseconds, whereas the delay time and slice timing is on the order of milliseconds. Thus, $T_1(x)$ is

$$\frac{2e^{-\frac{TI}{T_1(x)}}}{1+e^{-\frac{TI}{T_1(x)}}} = 1 - \frac{S_N(x)}{G_N} \tag{9}$$

After manipulation:

$$T_1(x) = -\frac{\mathrm{TI}}{\ln\left(\frac{S}{2-S}\right)} \tag{10}$$

Given a dataset with a T1-weighted MPRAGE, T2-weighted dual echo, and labels, these images together with equations (1), (4), (5), and (10) can be used as synthetic atlases.

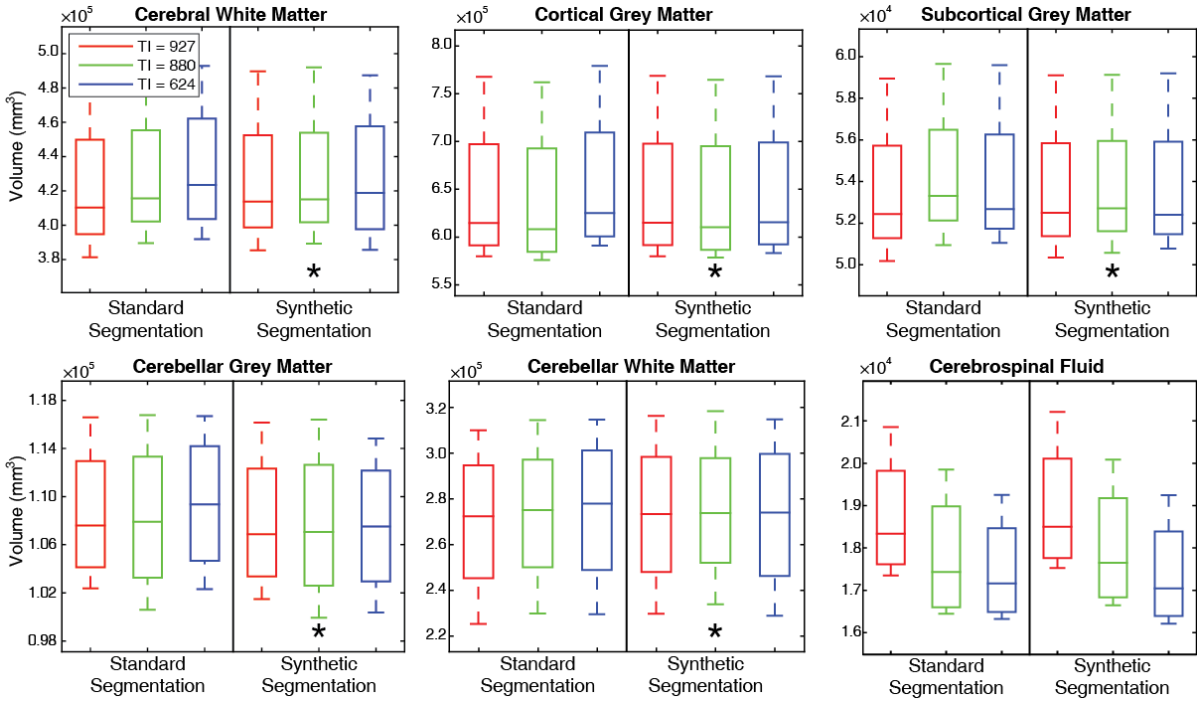

Figure III-3 Volumetric segmentation results for the standard and synthetic multi-atlas segmentation approaches on seven acquired subjects. For most of the major tissue types, there was a significant reduction in variance using the synthetic segmentation compared with the standard (*, $p<0.05$). Cerebrospinal fluid did not show a significant improvement using either method.

2.2. Synthetic Atlas Creation

The Kirby 21 Multi-Modal MRI Reproducibility Resource consists of 21 subjects where each subject was scanned consecutively with 12 anatomic and quantitative imaging sequences [125]. The Kirby 21 captured a T2-weighted dual echo (TR/TE1/TE2=6653ms/30ms/80ms) and T1-weighted MPRAGE (TR/TE/TI=6.7/3.1/842ms) so that the underlying T1- and T2-relaxation and proton density maps can be derived [125]. Of the 21 subjects available, the ten subjects whose white matter T2-relaxation matched the literature [124] were selected to be used as atlases. Lastly, the labels were derived using the publicly available BrainCOLOR atlases (www.neuromorphometrics.com) and a previously described multi-atlas segmentation procedure [92, 93, 97]. The result of this is a series of T1 relaxation, T2 relaxation, PD, and label volumes. Herein the T1 and T2 relaxation and PD volumes are the synthetic image set.

2.3. Synthetic Multi-Atlas Segmentation

The synthetic MAS procedure directly follows the standard MAS procedure[109], except the MR signal equations are used to create atlas images of similar intensity contrast to the target.

	Standard Segmentation			*Synthetic Segmentation*		
	TI = 927	TI = 880	TI = 624	TI = 927	TI = 880	TI = 624
Cerebral White Matter	4.14×10^5	4.19×10^5	4.21×10^5	4.18×10^5	4.18×10^5	4.20×10^5
Cortical Grey Matter	6.13×10^5	6.10×10^5	6.25×10^5	6.13×10^5	6.11×10^5	6.15×10^5
Subcortical Grey Matter	5.22×10^4	5.31×10^4	5.27×10^4	5.26×10^4	5.28×10^4	5.26×10^4
Cerebellar Grey Matter	1.07×10^5	1.08×10^5	1.09×10^5	1.07×10^5	1.07×10^5	1.08×10^5
Cerebellar White Matter	2.74×10^5	2.75×10^5	2.76×10^5	2.75×10^5	2.75×10^5	2.75×10^5
Cerebrospinal Fluid	1.83×10^4	1.74×10^4	1.71×10^4	1.84×10^4	1.76×10^4	1.70×10^4

Table III-1 Summary of average segmentation volumes for the three acquired scans on seven subjects. Average volumes for the two segmentation techniques, standard segmentation and synthetic segmentation, are shown.

The synthetic atlases are used for both the registration and segmentation. It is important to note that the labels do not change in the synthesis process.

3. Methods

Briefly, three distinct MR sequences were acquired for a cohort of seven healthy control subjects to assess the sequence variability under controlled circumstances. These scans were segmented using both the standard and synthetic multi-atlas segmentation approaches. The acquired data are segmented with a standard multi-atlas segmentation approach. The synthetic image set are used as target images where the sequence parameters were varied over a practical range of MPRAGE parameters. We show that the acquired target data and the synthetic target data show a similar pattern, implying that the biases based on the different sequences are captured by the synthetic data. Finally, the ABIDE study is considered, where autistic and healthy subjects are scanned at over 20 locations around the world. At these locations, different sequences are used, thus decreasing the consistency in the segmentation results.

3.1. Data Acquisition

Three distinct T1-weighted MPRAGE sequences were acquired on a cohort of seven control subjects (2F/5M, 21-58 years old, no history of neurological disorders). All scans were acquired without re-positioning in a one-hour session for each patient on a 3T Philips Achieva MRI (Philips Medical Systems, Best, The Netherlands) with a 32-channel receive coil. The first sequence had TI/TR/TE/α 624/8.9/4.6ms/8°, the second sequence had TI/TR/TE/α 891/8.2/3.7ms/8°, and the third sequence had TI/TR/TE/α 927/7.9/3.6ms/5°. Note that the sequences are representative of MPRAGE sequences used routinely at Vanderbilt University, and the second sequence is within the ADNI sequence parameter limits.

3.2. Synthetic Multi-Atlas Segmentation of Acquired Data

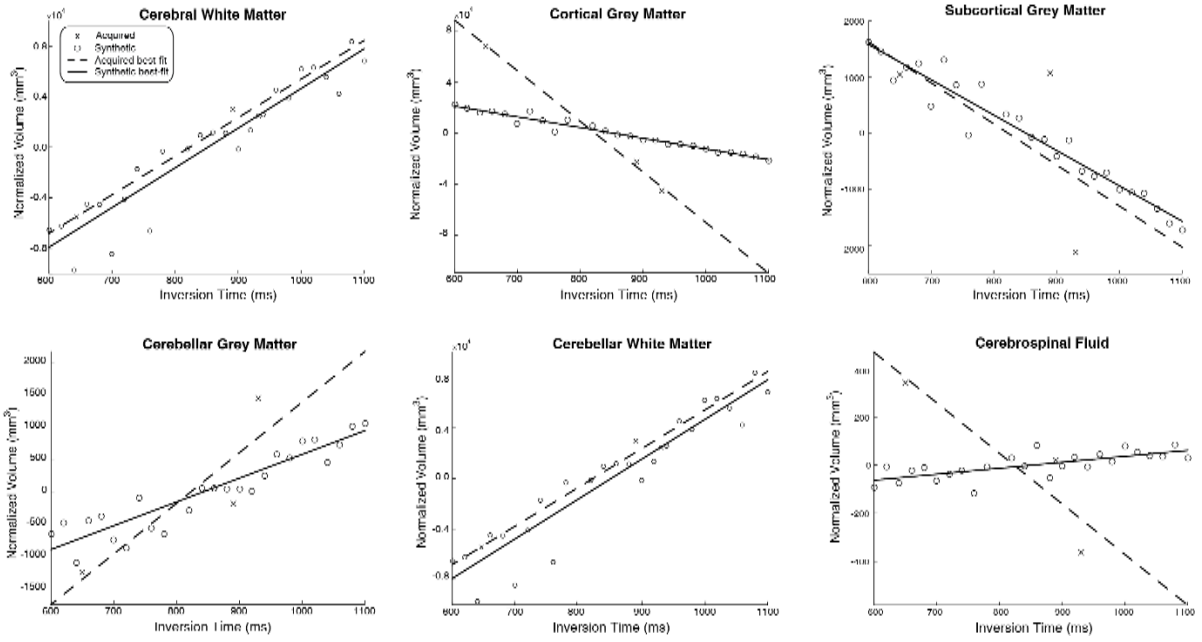

Figure III-4 Comparison of trends in acquired data compared with synthetic. The trend lines and points for one subject of each are shown here, normalized to a mean of zero. Six tissue types are presented here and cerebral white matter, cortical grey natter, subcortical grey matter, cerebellar grey matter, and cerebellar white matter showed similar trends between the synthetic and acquired images. Similar trends are present across the whole population and are statistically significantly similar between the acquired and synthetic data.

Two multi-atlas segmentations were performed for each of the 21 acquired scans. First, each scan was segmented using the *standard* multi-atlas pipeline outlined in [109]. For this segmentation, the ten Kirby-21 images selected in §2.2 were used as atlases. For this segmentation, the acquired T1-weighted MPRAGE was used as the atlas image volume and the labels derived in §2.2 for these ten images were used. Second, each scan was segmented using the *synthetic* multi-atlas pipeline outlined previously using the ten synthetic atlases derived in §2.2.

3.3. Standard Segmentation of the Acquired and Synthetic Data

To assess the sensitivity of standard multi-atlas segmentation to inversion time, target images were created using each of the ten synthetic atlas images with inversion times ranging from

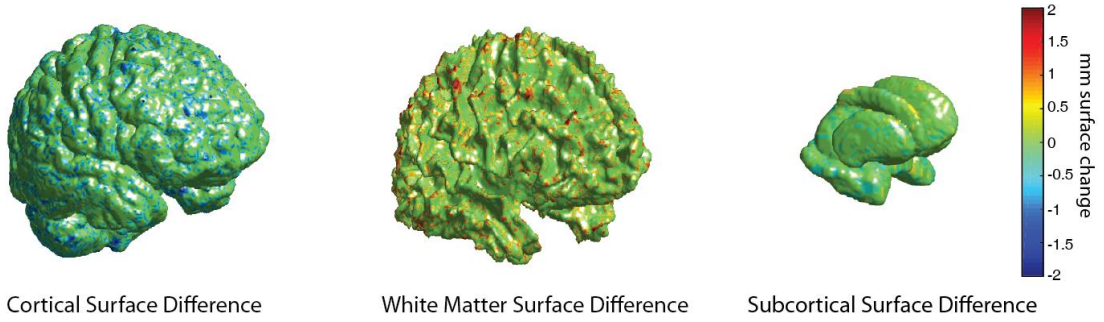

Figure III-5 Qualitative surface changes comparing the standard and synthetic multi-atlas segmentations for one randomly selected subject within ABIDE. Surface distance is measured as the distance between the surface of standard multi-atlas segmentation and the synthetic multi-atlas segmentation. Negative changes corresponds to cases where the synthetic segmentation's surface was within the standard. Synthetic segmentation tends to pull the grey matter/CSF boundary inward whereas the synthetic segmentation extrudes the white matter at similar sulci peaks. The subcortical changes are less consistent.

600 to 1100ms in intervals of 20ms. The standard multi-atlas segmentation pipeline [109] was performed on the 21 synthetic and one original image for each of the ten subjects. Segmentation volumes for these were then correlated with the true inversion time for the acquired data and the synthetic inversion time.

3.4. Synthetic Multi-Atlas Segmentation of ABIDE

The Autism Brain Imaging Exchange (ABIDE) provides multi-site, multi-modal brain imaging for autistic patients and healthy controls [118]. Each subject was scanned with a T1-weighted sequence and an fMRI, but each site designed its own sequences. No two sites share a sequence, which results in a significant site-effect for volumetric differences as has been reported in [113, 114]. We performed the standard MAS and synthetic MAS techniques on each ABIDE subject, matching the synthetic sequence information to where the scan was performed. A cross-validated classification of autistic versus healthy was performed where each site was held out. An L1-normalized logistic regression model was built including age, age squared, gender, gender

cross age, gender cross age squared, and the volumes from the 132 regions of interest to predict autism diagnosis status [126]. This procedure was performed twice, once using the synthetic MAS segmentation volumetric results and once using the standard MAS segmentation volumetric results. The "leave-site-out" procedure examines if the learned classifier performs better on data from sequences it was not familiar with, since no two sites shared a sequence.

4. Results

For the seven subjects acquired with three distinct MPRAGE sequences (§3.2), the 132 regions from the BrainCOLOR were condensed to six biologic types: cortical grey matter, cerebral white matter, subcortical grey matter, cerebrospinal fluid, cerebellar grey matter, and cerebellar white matter. The 132 regions were condensed to these because they are tissue groups with similar T1-relaxation and proton density profiles [124]. For both segmentation techniques, the variance of segmentation volume between the three sequences was calculated for each of the six structures. That is to say, for a given structure of interest (i.e. cortical grey matter) and a segmentation technique (i.e. synthetic multi-atlas segmentation) the variance in volume was calculated between the segmentations of the three T1-weighted imaging sequences. There was a significant decrease in the variance between the segmented volumes of total cortical grey matter, total cerebral white

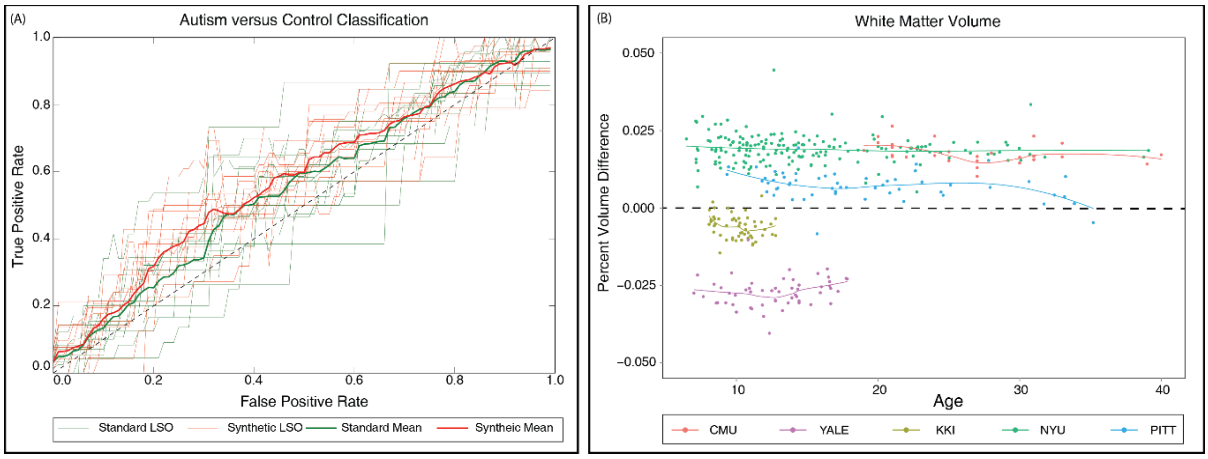

Figure III-6 In the leave-site-out classification (A) of autism versus control, the mean AUC was significantly higher using the synthetic segmentation volumes compared with the standard segmentations. Consistent trends in percent volume difference between the standard and synthetic segmentations were seen for these subjects (B). Site was significantly correlated with percent volume difference ($p<0.01$, Pearson's correlation) whereas age was not.

matter, and total subcortical grey matter volume ($p<0.01$, Wilcox sign-rank test; Figure 3; table 1). Cerebrospinal fluid showed no significant difference in variance using either method. Across five major tissue types of interest, cortical grey matter, cortical white matter, subcortical grey matter, cerebrospinal fluid, cerebellar grey matter, and cerebellar white matter, the mean effect of sequence on volume was 9% using the standard MAS. Using the synthetic MAS, this mean effect dropped to 4%. Qualitatively, these segmentations show consistent surface differences (Figure 4). Surface changes are localized primarily at the gyri for the cortical grey matter and cerebral white matter and on a regional basis for the subcortical grey matter.

The standard multi-atlas segmentations performed on the synthetic target images and the acquired images with varying inversion times showed similar trends. The trends of cortical grey matter, cortical white matter, and subcortical grey matter showed significantly similar patterns between the acquired and synthetic target images (Figure 5; $p<0.05$, Wilcox rank-sum test).

However, cerebrospinal fluid did not show a significant correlation between these acquired and synthetic data.

In the autistic versus healthy classification, there was an improvement in classification using the synthetic multi-atlas segmentation compared with the standard multi-atlas segmentation. There was a significant improvement in area under the receiver operator characteristic (AUC) curve in the leave-site-out classification ($p < 0.01$, Wilcox sign-rank test; Figure 6 A). Mean AUC increased from 0.57 (standard deviation 0.03) to 0.61 (standard deviation 0.03), and these results are comparable to previous studies [118]. Of the 17 ABIDE sites considered, 14 showed an increase in AUC using the synthetic multi-atlas segmentation. There was also no significant effect shown between age, race, or gender with percent volume change for any region of interest (Figure 6 B)

5. Discussion and Conclusions

Multi-atlas segmentation is currently one of the leading techniques for volumetric segmentation, and it is not without the sequence bias. By synthesizing atlases similar to [119], some of these sequence effects can be mitigated. Using synthetic atlases which match the sequence information of the scans, provides the atlases with information more consistent with the target images, thus decreasing some of the variance and improving statistical power when a linear effect cannot be used.

Imaging sequence plays a significant role in volumetric segmentation results, and thus segmentation is not an absolute quantitative process when working between sequences. This result has been reported previously with FreeSurfer, where a linear effect model was used to mitigate adverse effects [114]. We observe a primarily linear effect with inversion time, which allows linear effects modeling. Note that not all study designs permit modeling the sequence effect separately,

for instance if the control population for a study was provided separately from the patient population and the two populations had different sequences.

The synthetic MAS procedure presented here has several limitations. First, the labels used in our presented segmentation where automatically derived through a separate MAS procedure. As a result, any bias from either the sequences or the MAS technique present in that original procedure is propagated to these atlases and further to the target images. Second, the imaging used were not in the same space and were not at the same resolution. Since much of the sequence effects are in the partial voluming, not having comparable sequences causes some of the necessary information for the synthetic atlases to be lost in interpolation and different voxel sizes. Third, the T2-weighted sequence suffered from the first relaxation times being in the steep part of the exponential decay curve, and thus significant bias from that measure. Even with these biases, the synthetic multi-atlas segmentation still showed significant improvements in the volumetric segmentations.

Chapter IV

Automated, Open-Source Segmentation of the Hippocampus and Amygdala with the Open Vanderbilt Archive of the Temporal Lobe

1. Introduction

The hippocampus and amygdala are widely studied medial temporal lobe structures critical for memory and emotion, respectively [127, 128]. The hippocampus has been implicated as an important structure in the pathophysiology of Alzheimer's disease, schizophrenia, depression and epilepsy [129]. Rather than being a unitary structure, the hippocampus can be divided into subregions along its transverse and longitudinal axes [130, 131]. Subfields defined within the transverse axis, including CA fields 1-4 and the dentate gyrus, contribute to distinct memory functions [130, 132-134]. Anterior and posterior subregions differ in structural connectivity, function, and gene expression [135-137]. The amygdala is tightly connected to the hippocampus and plays an important role in regulating interpretation of facial expression, fear processing, and emotional learning [138]. Structural and functional changes in the amygdala have been identified in neuropsychiatric disorders including autism, anxiety, and schizophrenia [12, 139-144].

Advances in magnetic resonance imaging (MRI) allow for high-contrast, reproducible methods for visualization of the amygdala and hippocampus on standard T1-weighted images [145]. Traditionally, volumetric studies of these regions have been done by labor-intensive manual segmentation. The advancement of large-scale and longitudinal imaging studies in recent years necessitates the development of low-cost, reliable segmentation methods. Several open-source, automatic techniques have been developed for segmentation of the hippocampus and amygdala using standard, T1-weighted MR images. One of the most common techniques, FreeSurfer,

reconstructs the cortical surface and several subcortical brain structures with an energy model [72]. Another common technique is FSL FIRST, which uses a Bayesian shape and appearance model to segment subcortical structures [146]. A third approach is a multi-atlas segmentation [93, 109]. Other approaches, such as Automated Segmentation of Hippocampal Subfields (ASHS) delineate the hippocampus and its subfields, but require collection of an additional MR sequence outside of a standard T2-weighted MR protocol [147].

Multi-atlas segmentation (MAS) can provide a robust and accurate segmentation of a target image [109]. Previous works have used MAS to segment the whole brain focusing on the cortex, the optic nerve, the abdomen, and other structures [93, 148-150]. A typical multi-atlas segmentation procedure involves non-rigidly registering ten or more atlases, image volumes paired with expertly delineated labels, to a target image to be segmented. Note that atlases are image sets with one or more structural images and a set of labels corresponding to the structures of interest. These registered target images are then joined together to create a representation that is more accurate than any individual registered atlas. One typical assumption of many studies is that 30 atlases is sufficient to produce a segmentation to produce a maximally accurate segmentation [109].

In this work, we present the Open Vanderbilt Archive of the temporal Lobe (OVAL). OVAL is a fully automated segmentation approach using 195 atlases to produce an accurate segmentation of the hippocampus head, body, and tail and the amygdala. OVAL uses a previous multi-atlas segmentation of the whole brain, a common practice in most neuroimaging techniques, to localize the hippocampus and amygdala. OVAL then registers the 195 atlases to the localized target images and fuses them following a standard MAS protocol. OVAL produces segmentations of the amygdala and hippocampus more accurate than other common open-source tools and

produces segmentations of the hippocampus head and body comparably accurate with expert reproducibility. Moreover, OVAL allows us to test the assumption that 30 atlases is enough for optimal multi-atlas segmentation, and we show that 30 atlases produce inferior results to using the entire population of 195 atlases.

2. Methods

2.1. Overview

The OVAL algorithm produces segmentations of target images using 195 atlases of the hippocampus and amygdala. The atlases are generated from 195 manually delineated hippocampi (dataset 1) and automatically segmented amygdalae defined from training data in a second population of 35 subjects with manually delineated amygdalae (dataset 2). Briefly, the 195 subjects with manual hippocampus segmentations were segmented with the 35 amygdala atlases following the protocol outlined below. These atlases are then cropped to a bounding box around each temporal lobe, resulting in 195 left and 195 right atlases. For a given target image, the atlases are used in a MAS framework to segment the amygdala, hippocampus head, and body. Finally, an anatomical landmark defined from whole-brain segmentation is used to split the hippocampus body into the body and tail.

2.2. Subjects and Image Acquisition

Dataset 1 consisted of MR images acquired in 90 healthy adults and 105 adults with a non-affective psychotic disorder (56 schizophrenia; 32 schizoaffective disorder; 17 schizophreniform disorder) taken from an ongoing study of psychiatric phenotypes (table 1). Patients were recruited from the Vanderbilt Psychotic Disorder Program and controls were recruited from the surrounding Vanderbilt community. All participants were assessed with the Structured Clinical Interview for

	Dataset 1: Hippocampus		Dataset 2: Amygdala	
	Psychosis	**Control**	**Patient**	**Control**
N	105	90	18	17
Age, years (Mean ± SD)	34.62 ± 12.38	33 ± 11.33	24.6 ± 5.1	23.6 ± 4.8
Gender (Female/Male)	37/68	41/49	12/6	10/7
Race (White/Black/Other)	63/37/5	60/26/4	13/1/3	15/2/0

DSM-IV [151]. New York: Biometrics Research, New York State Psychiatric Institute (2002). Dataset 2 included 35 subjects recruited as part of a study on temperament. The Vanderbilt University Institutional Review Board approved both studies. Structural images were acquired with a 3D T1-weighted MPRAGE sequence (TI/TR/TE = 860/8.0/3.7 ms; 170 sagittal slices; voxel size = 1.0mm^3). All images were collected on a Philips Achieva scanner (Philips Healthcare, Inc., Best, The Netherlands).

2.3. Hippocampus Protocol

Manual delineation of the hippocampus on images from dataset 1 was completed following a previously published protocol [152, 153]. For the purposes of this study, the term hippocampus includes the hippocampus proper (CA1-4 and dentate gyrus) and parts of the subiculum, together more often termed the hippocampal formation [154]. Briefly, the hippocampus was traced as follows. Beginning with the right hippocampus, the full structure was traced from lateral to medial. The tracing was then re-examined in the coronal plane and refined to be consistent in both planes.

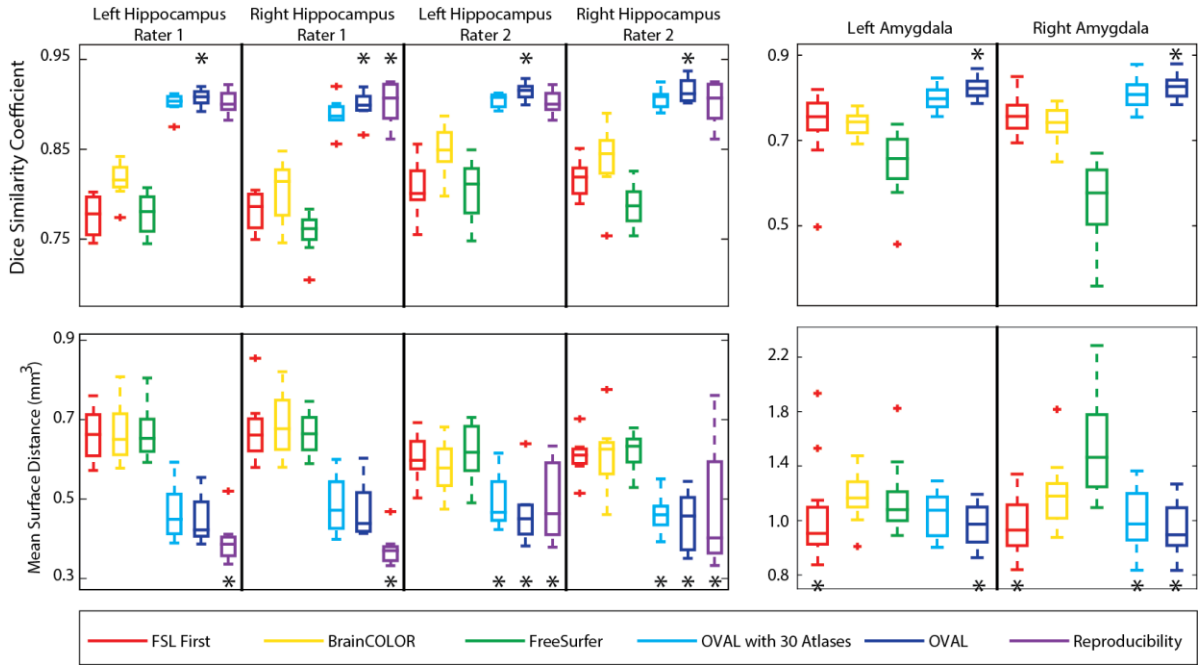

Figure IV-1: Quantitative segmentation results for the whole hippocampus and amygdala. OVAL outperforms all other segmentation techniques in terms of DSC for the left hippocampus in both raters, the right hippocampus in rater 2, and the left and right amygdala ($p<0.05$, *). OVAL outperforms all other techniques for the right hippocampus of rater 1 except human reproducibility, which performs comparably statistically. Human reproducibility outperforms all other techniques in MSD for the left and right hippocampus for rater 1 ($p<0.05$, *). OVAL and OVAL with 30 atlases outperform all other automated techniques for those structures. OVAL, OVAL with 30 atlases, and human reproducibility perform statistically comparable for the left and right hippocampus for rater 2 and outperform all other techniques ($p<0.05$, *). OVAL and FSL FIRST perform statistically similarly for the left amygdala and outperform all other segmentation approaches ($p<0.05$, *). OVAL, OVAL with 30 atlases, and FSL FIRST perform statistically similarly for the right amygdala and outperform all other segmentation approaches ($p<0.05$, *).

The anterior and posterior regions were divided in the coronal plane at the slice where only one cut through the hippocampus remained visible. The tracing was then verified in the sagittal view. This process was then repeated for the left hippocampus. The resulting 195 labeled images are hereafter referred to as hippocampus atlases.

2.4. Amygdala Protocol

Manual delineation of the amygdala on images from dataset 2 was completed as described in Clauss et al., 2014. Structural images were first normalized to Montreal Neurologic Institute (MNI) standard template space. The amygdala was then traced in the axial plane, proceeding from superior to inferior and boundaries were refined in the coronal and sagittal plains. The 35 labeled images from dataset 2 are hereafter referred to as amygdala atlases.

2.5. Whole-Brain Segmentation

Whole-brain segmentation (WBS) was carried out on target images from dataset 1. First, 45 atlases labeled with the BrainCOLOR protocol (www.neuromorphometrics.com) were affinely registered to each target image with Niftyreg [86]. The 15 atlases geodesically closest to the target were selected and these atlases were non-rigidly registered to the target image using the Advanced Normalization Tools (ANTs) Symmetric Normalization (SyN) algorithm [87]. The 15 registered atlases were fused with the hierarchical Non-Local Spatial STAPLE algorithm [92, 93, 97]. Finally, the segmentation was refined with corrected learning following [99]. The resulting segmentation contained 132 labels including the hippocampus and amygdala in both hemispheres, along with 98 other cortical structures. This segmentation acts as a guiding mechanism for segmentation of the hippocampus and amygdala.

2.6. Atlas Creation

Initially, two populations of data were available. The first population consisted of 195 subjects (90 healthy adults, 105 adults with Schizophrenia) scanned with a T1-weighted MPRAGE scan (TI/TR/TE=860/8.0/3.7ms) and manually labeled with the protocol from §2.1, herein *hippocampus atlases*. The second population consisted of 35 subjects (15 healthy, 20 adults with

Schizophrenia) scanned with a T1-weighted MPRAGE scan (TI/TR/TE=860/8.0/3.7ms) and manually labeled with the protocol from §2.2, herein *amygdala atlases*. To create a set of atlases labeled with both the amygdala and hippocampus, the 35 amygdala atlases were non-rigidly registered to the hippocampus atlases and the registered atlases were fused with joint label fusion (JLF). The resulting amygdala labels were added to the hippocampus atlases where they did not conflict with manual hippocampus labels.

The hippocampus atlases were then segmented with the WBS described in §2.3. For each atlas, the WBS was used to determine a bounding box around the hippocampus and amygdala for each hemisphere; the bounding box was dilated 5mm in each direction to assure the full true hippocampus and amygdala was included. The bounding box was then used to extract the atlas image and label volume localized to the region around the hippocampus and amygdala. This resulted in 195 hippocampus and amygdala atlases for each hemisphere, herein *temporal lobe atlases*.

2.7. OVAL Segmentation

OVAL segmentation results in lateralized segmentations of the amygdala, hippocampus head, body, and tail. The algorithm requires an input T1-weighted MRI volume and a WBS. First, the input T1-weighted volume is cropped to the left hippocampus and amygdala by its WBS, herein the *left target image*. The 195 left temporal atlases are non-rigidly registered to the left target image with NiftyReg and the ANTs SyN algorithm [86, 87]. The atlases are fused with JLF and the posterior probability volumes for amygdala, hippocampus head, and body are considered [98]. At

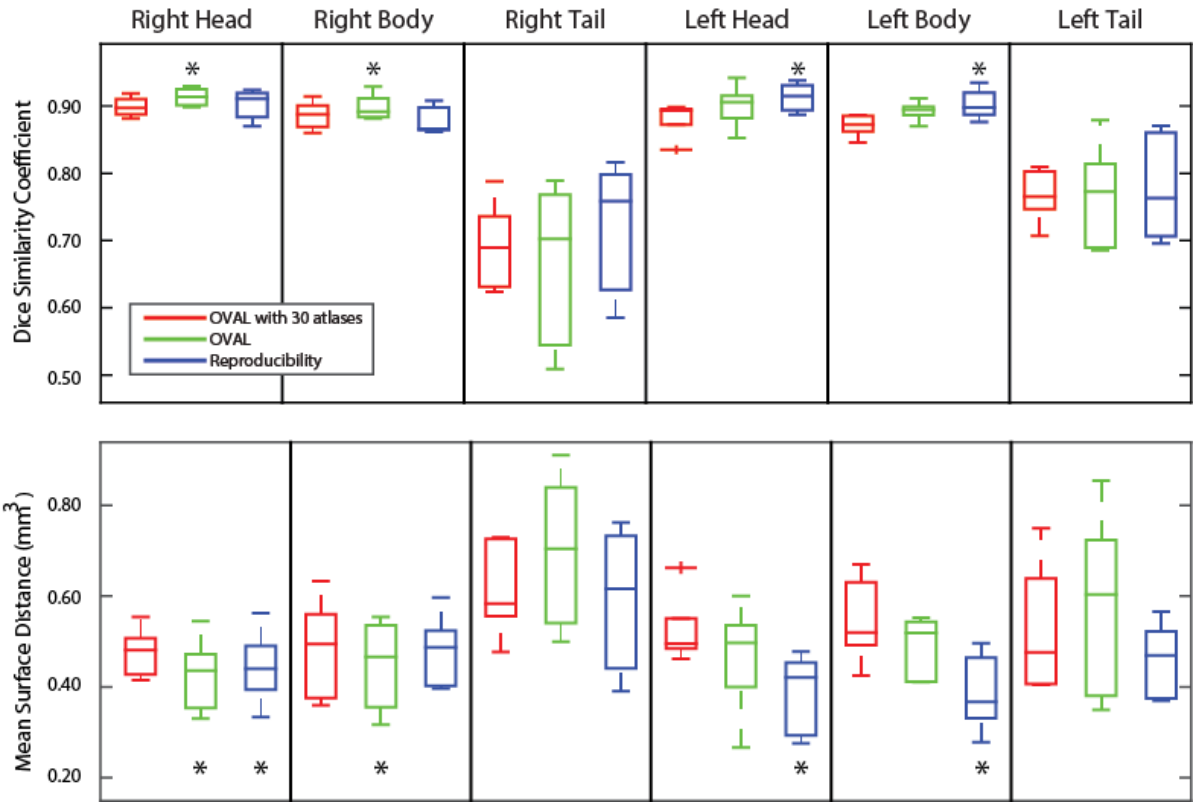

Figure IV-2 Quantitative segmentation results for the whole hippocampus head, body, and tail. OVAL outperformed OVAL with 30 atlases and human reproducibility on the right head and body in Dice Similarity Coefficient ($p<0.05$; *). No technique showed significant improvement on the right or left tail. Human reproducibility outperformed OVAL and OVAL with 30 atlases ($p<0.05$, *), though OVAL outperformed OVAL with 30 atlases. In mean surface distance, OVAL and human reproducibility outperformed OVAL with 30 atlases for the right head, OVAL outperformed all other techniques for the right body, no technique outperformed any other for the right and left tail, and human reproducibility outperformed the other techniques for the left and right head ($p<0.05$, *).

voxels where the sum of the probability of these labels exceeds 0.5, the label of these three with the highest probability is chosen

$$L_i = \begin{cases} \arg_{X=S} \max p_i^X & p_i^{GM} > 0.5 \\ \text{background} & p_i^{GM} \leq 0.5 \end{cases} \quad 16$$

where L_i is the label decision at voxel i, p_i^{GM} is the sum of the probability of amygdala, hippocampus head, and body at I, p_i^X is the probability of label X at i, and S is the set of labels of interest, amygdala, hippocampus head, and hippocampus body. This correction primarily applies

to voxels near the boundary of two structures, for instance the hippocampus head and body, where JLF shows a posterior probability less than 0.5 for the background, but the probability of the head or body does not exceed the probability of background. For instance, a case where the probability of hippocampus head is 0.35, hippocampus body is 0.25, and background is 0.4. This procedure is then repeated for the right hippocampus.

After the segmentation of the amygdala, hippocampus head, and hippocampus body is complete, the final step in the segmentation is splitting the body and tail. For the left hippocampus, the most posterior point on the left thalamus is identified from the WBS by finding the point on the thalamus nearest to the mean location of the left occipital lobe. Next, a line is fit through the coordinates of the voxels of the full hippocampus, defined by the OVAL segmentation. Lastly, a plane is fit through the posterior point of the thalamus, orthogonal to the line through the hippocampus. The points of the body of the hippocampus posterior to the plane are then defined as the tail and the points anterior to the plane defined as the body.

3. Results

Three experiments were considered to test the accuracy of OVAL compared with other segmentation approaches. A set of 10 atlases, distinct from the training population, was labeled with the hippocampus segmentation protocol (§2.2) by two expert raters, creating a human reproducibility dataset, herein the hippocampus testing atlases. A separate set of 35 atlases, distinct from the training population and the hippocampus reproducibility population, was labeled with the amygdala protocol (§2.3), herein the amygdala testing atlases. These two datasets were segmented

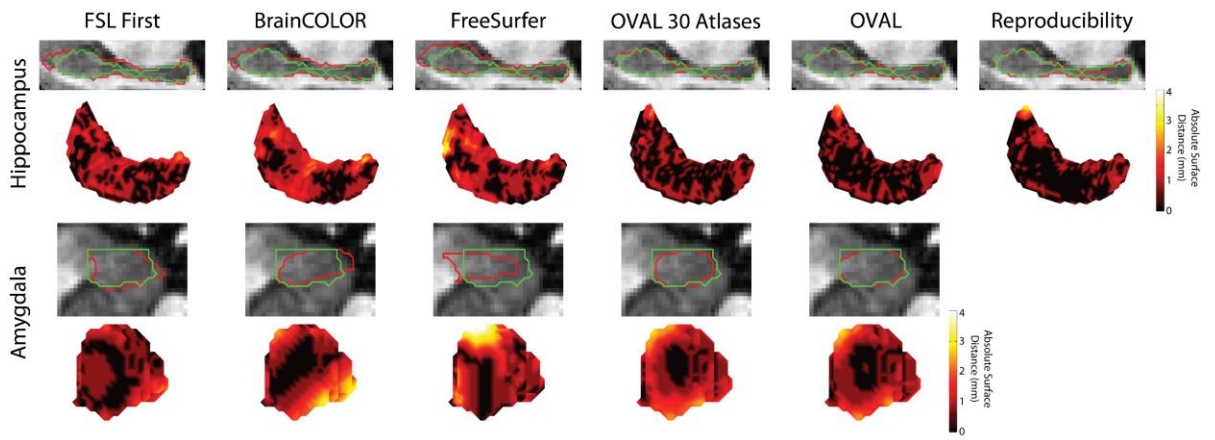

Figure IV-3 Median qualitative segmentation results for the whole hippocampus and amygdala; red represents the estimated segmentation and green is the truth. FSL FIRST, BrainCOLOR, and FreeSurfer all showed large surface distances up to 4mm for both the hippocampus and amygdala. OVAL and OVAL with 30 atlases were typically within 1mm distance on the hippocampus, though OVAL produced more consistent results than OVAL with 30 atlases. On the amygdala, OVAL and OVAL with 30 atlases captured the overall contour of the amygdala, but were not able to accurately localize the borders since they are defined by anatomical landmarks.

with the WBS described in (§2.4), FreeSurfer, FSL FIRST, OVAL with 30 atlases, and OVAL with 195 atlases to test OVAL's accuracy. Next, the 10 hippocampus testing atlases were segmented with OVAL with 30 atlases and OVAL with 195 atlases to test OVAL's accuracy on the hippocampus head, body, and tail. Finally, since there is a significant amount of discrepancy between hippocampus segmentation protocols, the Kirby21 multi-modal reproducibility dataset was segmented with WBS in (§2.4), FreeSurfer, FSL FIRST, and OVAL to compare OVAL's intra-subject reproducibility of the amygdala and full hippocampus segmentation.

3.1. Whole Hippocampus and Amygdala Segmentation

The hippocampus testing atlases and amygdala testing atlases were first segmented with the WBS, identified as BrainCOLOR in figures and results. These atlases were segmented with FreeSurfer using their standard reconstruction and FSL FIRST with standard parameters. Finally,

the atlases were segmented with OVAL and OVAL with a subset of 30 atlases. The OVAL hippocampus segmentations were reduced to whole hippocampus by collapsing hippocampus head, body, and tail into one label. Mean surface distance (MSD) and Dice similarity coefficient (DSC) were calculated between each segmentation and the hippocampus and amygdala testing atlases.

For DSC of the hippocampus testing atlases, OVAL outperformed FSL FIRST, FreeSurfer, BrainCOLOR, OVAL with 30 atlases, and human reproducibility for left hippocampus for rater 1, left hippocampus rater 2, and right hippocampus rater 2 ($p<0.05$ Wilcoxon sign-rank test; figure 1). For right hippocampus rater 1, OVAL and human reproducibility performed comparably and outperformed other techniques ($p<0.05$ Wilcoxon sign-rank test). For MSD of the hippocampus testing atlases, human reproducibility outperformed other techniques for the left and right hippocampus for rater 1, OVAL with 30 atlases and OVAL outperformed all other automated techniques ($p<0.05$ Wilcoxon sign-rank test; figure 1). For the left and right hippocampus for rater 2, OVAL with 30 atlases, OVAL, and human reproducibility all performed statistically similarly and outperformed all other techniques ($p<0.05$ Wilcoxon sign-rank test; figure 1).

For DSC of the amygdala testing atlases, OVAL outperformed FSL FIRST, FreeSurfer, BrainCOLOR, and OVAL with 30 atlases for both left and right amygdala ($p<0.05$ Wilcoxon sign-rank test). For MSD of the amygdala testing atlases, OVAL and FSL FIRST outperformed BrainCOLOR, FreeSurfer, and OVAL with 30 atlases for the left amygdala ($p<0.05$ Wilcoxon sign-rank test; figure 1). For the right amygdala, OVAL, OVAL with 30 atlases, and FSL FIRST outperformed BrainCOLOR and FreeSurfer, but were not statistically separable ($p<0.05$ Wilcoxon sign-rank test; figure 1). For MSD of the amygdala, FSL FIRST and OVAL resulted in a significantly lower MSD for the left amygdala compared with FreeSurfer, BrainCOLOR, and

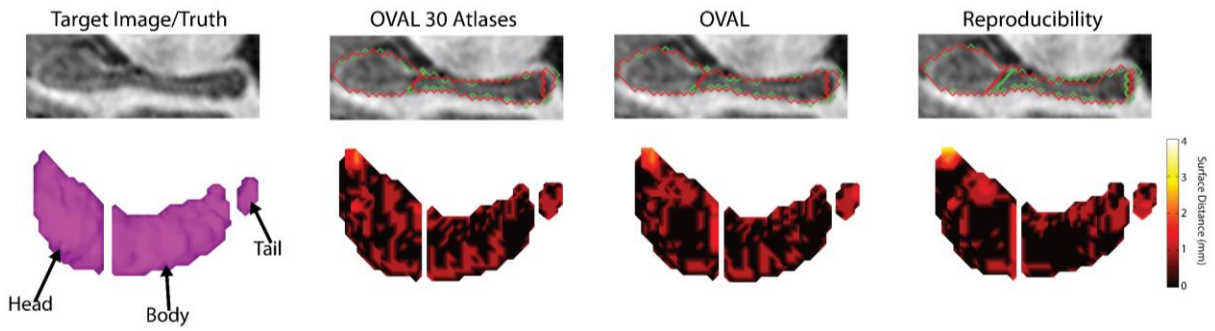

Figure IV-4 Median qualitative segmentation results for the hippocampus head, body, and tail. Green represents the true segmentation and red represents the estimate. Human reproducibility defined a different point for the head/body split and rater 2 under-segmented the tail of the hippocampus and the tip of the head compared with rater 1. OVAL with 30 atlases produced more local errors than OVAL. Images were rotated along the axis of the hippocampus, gaps between the head, body, and tail are exaggerated for visualization.

OVAL with 30 atlases; for the right amygdala, FSL FIRST, OVAL with 30 atlases, and OVAL resulted in a significantly lower MSD than BrainCOLOR and FreeSurfer ($p<0.05$; figure 1).

3.2. Hippocampus Head, Body, Tail Segmentation

The hippocampus testing atlases were segmented with OVAL, following §2.6, and OVAL with 30 atlases. Since no other approach provides a segmentation of the hippocampus head, body, and tail, only these two approaches were compared against reproducibility. Since the hippocampus testing atlases only segmented the head and body, the tail was split from the body following the protocol in §2.6. For simplicity, only the results with respect to rater 1 are presented (figure 2).

For the right head and body, OVAL significantly outperformed human reproducibility and OVAL with 30 atlases in DSC ($p<0.05$ Wilcoxon sign-rank test). OVAL significantly outperformed human reproducibility and OVAL with 30 atlases in MSD of the right head and OVAL and human reproducibility outperformed OVAL with 30 atlases in MSD of the right tail

($p<0.05$ Wilcoxon sign-rank test). For the left head and body, human reproducibility outperformed OVAL and OVAL with 30 atlases in DSC and MSD, though OVAL outperformed OVAL with 30 atlases on both of these structures ($p<0.05$ Wilcoxon sign-rank test). For the right and left tail, no technique performed significantly better in DSC and MSD.

3.3. Whole Hippocampus and Amygdala Reproducibility

The BrainCOLOR, FreeSurfer, and FSL FIRST segmentation approaches considered in §3.1 do not use the same protocol as the manual segmentations [155]. Thus, saying that the OVAL approach has higher DSC and lower MSD than the other approaches does not necessarily conclude that it is a better approach than the other techniques. The Kirby21 multi-modal reproducibility dataset is a set of 21 subjects scanned twice in immediate succession. The two T1-weighted MPRAGEs for each subject was segmented with the BrainCOLOR multi-atlas segmentation (§2.4), FreeSurfer, FSL FIRST, and OVAL. To assess the reproducibility of each technique, the volume of the amygdala and whole hippocampus was calculated. The average volume ($AV = \frac{\text{volume}_1+\text{volume}_2}{2}$) and absolute percent volume difference ($PVD = \frac{|\text{volume}_1+\text{volume}_2|}{AV}$) between each scanning session was calculated for each subject. The percent volume difference of OVAL was significantly lower than all other techniques for all structures ($p<0.05$ Wilcoxon sign-rank test; figure 3). In the hippocampus, OVAL had an average percent volume similarity of 0.75 and 0.66 for the left and right respectively and for the amygdala, OVAL had an average percent volume similarity of 2.67 and 3.10 for the left and right respectively.

4. Discussion

In this work, we presented the OVAL segmentation algorithm for segmentation of the hippocampus and amygdala. First, we presented labeling protocols for the hippocampus head and

body and the amygdala. Second, we created an atlas population of 195 subjects with manually traced hippocampi and automatically segmented amygdalae. Third, we presented the OVAL segmentation algorithm which, for a given target image, uses an initialization of the temporal lobe from a whole-brain segmentation to efficiently perform the segmentation.

OVAL was evaluated in three experiments. First, OVAL was compared with FreeSurfer, FSL FIRST, a whole-brain multi-atlas segmentation with BrainCOLOR, and OVAL with a subset of 30 atlases for segmentation of the whole hippocampus and amygdala. OVAL outperformed all other approaches for segmentation of the hippocampus and amygdala. Qualitatively, the OVAL segmentation tends to produce a segmentation of the amygdala with smoother boundaries than the atlas definitions since several of the atlas boundaries are defined by global landmarks instead of boundaries visible in contrast (figure 4). Second, OVAL was evaluated against human reproducibility and OVAL with a subset of 30 atlases for segmentation of the hippocampus head, body, and tail. In general, OVAL performed comparably with human reproducibility and outperformed OVAL with 30 atlases. Qualitatively, OVAL segmentations are typically within 1mm of the truth at all voxels including the boundary between the body and head (figure 5). Third, since FreeSurfer, FSL FIRST, BrainCOLOR, and OVAL use different segmentation protocols, these segmentation protocols were evaluated for reproducibility with the Kirby21 multi-modal reproducibility dataset. OVAL showed the lowest average percent volume similarity of any technique, implying that it is the most reproducible of any algorithm tested.

OVAL presents an accurate and reproducible segmentation of the hippocampus and amygdala, two of the most studied structures in the human brain. Furthermore, since OVAL has 195 atlases available, it was possible to test whether 30 atlases were sufficient to produce optimal segmentations or if the full atlas population produced better results. OVAL using all atlases

outperformed OVAL using only 30 atlases, proving that 30 atlases is not sufficient to produce optimal segmentations for this task.

Chapter V

Improving Cerebellar Segmentation with Statistical Fusion

1. Introduction

The cerebellum is an anatomic region of the central nervous system located in the posterior fossa, inferior to the cerebrum and posterior to the brain stem. As with the cerebrum, the cerebellum consists of two hemispheres (left and right), but also contains midline gray matter structure known as the vermis [156-158]. The cerebellum consists of a layer of tightly folded gray matter surrounding densely packed white matter beneath. The white matter contains four gray matter nuclei: the dentate, globose, emboliform, and fastigial, which receive input fibers from the cerebellar cortex and output to the cerebrum; these cerebellar nuclei account for most of the fibers leaving the cerebellum. The somatotopically organized cerebellum plays an important role in motor function and secondary roles in higher order cognition and decision making. Segmentation of the cerebellum provides a unique challenge in that the cerebellar lobules are not easily differentiated in healthy subjects due to the resolution of the imaging whereas subjects with cerebellar atrophy have more easily differentiable structures (Figure 1).

Automated segmentation of the cerebellum has been deeply discussed and characterized in the literature. Van der Lijn used atlas registration and local feature descriptors to segment the left and right hemispheres of the cerebellum but did not segment any of the individual lobules or the vermis [159]. Saeed and Puri developed a semi-automated procedure using template selection and local texture to segment the whole cerebellum [160]. Powell et al use machine learning with

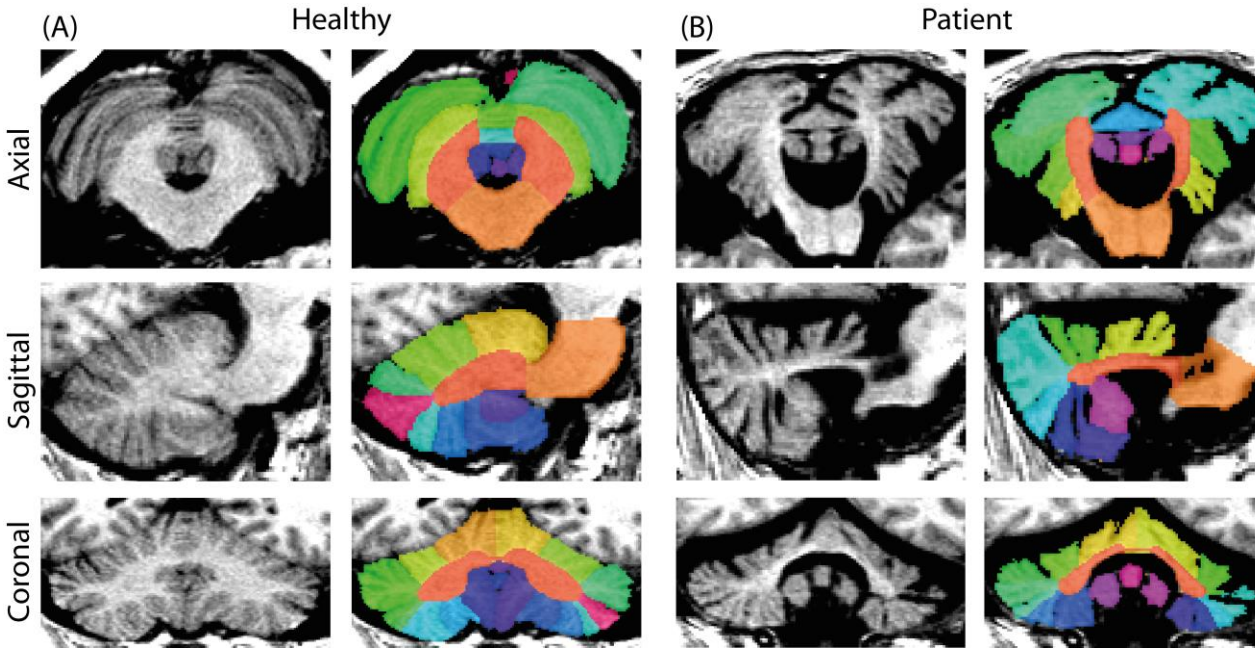

Figure V-1: Axial, coronal, and sagittal segmentation results for a healthy (A) subject and a patient with severe cerebellar ataxia (B). Note the easily differentiable lobules in the patient whereas the differentiation of the lobules is lost to the resolution of the imaging in the healthy subject.

probabilistic atlases to segment the cerebellum into upper, middle, and lower lobules but do not explore deeper characterization of regions and only apply their method to healthy subjects [161]. Diedrichsen et al present a probabilistic atlas for segmentation characterizing all of the cerebellar lobules but the single probabilistic atlas does not individually provide robust segmentations across diverse subject populations [162]. Lastly, Yang et al propose performing multi-atlas segmentation of the cerebellar lobules and vermis followed by a post-hoc graph cut to model the boundaries [101].

Herein we propose new segmentation algorithms which combines the ideas of patch-based correspondence of Coupe et al and strong internal atlas selection of Langerak and SIMPLE [96, 105]. The first algorithms, Local SIMPLE and Local Spatial SIMPLE, incorporate intensity information into the generative model of SIMPLE similar to the models of locally-weighted vote and we extend the model to allow spatially varying performance parameters [84, 163]. The third

algorithm we propose, Non-Local SIMPLE, combines the ideas of patch-based fusion with the strong semi-parametric atlas selection of SIMPLE, but instead of treating atlases independently, Non-Local SIMPLE assumes an independence between local patches and develops a performance model around them. We evaluate the effectiveness of these models on two distinct populations of cerebellum atlases and compare these algorithms to previous segmentation techniques.

2. Methods

We begin be defining the data and the standard pipeline used for multi-atlas segmentation. We then define the generative models underlying Local SIMPLE and Non-Local SIMPLE. For these models, we define: T_i is the true label at voxel i, s is an arbitrary label, I_i is the intensity observed at voxel i by the target image, D_i is a $1 \times R$ vector of labels observed at i, R is the number of available raters, L is number of observed labels, N is the number of observed voxels, b is an integer pooling region, A_i is a $1 \times R$ vector of the intensity values observed at i, c is a $1 \times R$ binary vector indicating the current atlas selection state for each atlas, ϵ is a $1 \times R$ rater error vector, σ is the standard deviation used in intensity weighting, k is the current iteration during expectation maximization, j is a particular rater, θ is a $R \times 2 \times L \times L$ matrix where $\theta_{jnss'}$ is the likelihood rater j observes s given that the true label is s' and their atlas selection status is n. For brevity the definitions of θ are left to Xu et al [163].

2.1. Data

Two datasets were considered in this study. The first dataset, herein Anura, consisted of 25 subjects, 13 with cerebellar ataxia and 12 healthy, ranging in age from 36 to 73, 23 female and 2 male, scanned with a 1.5T three dimensional SPGR sequence and cerebellum manually traced by a trained expert. The second dataset, herein AT, consisted of 45 subjects, 15 healthy controls and

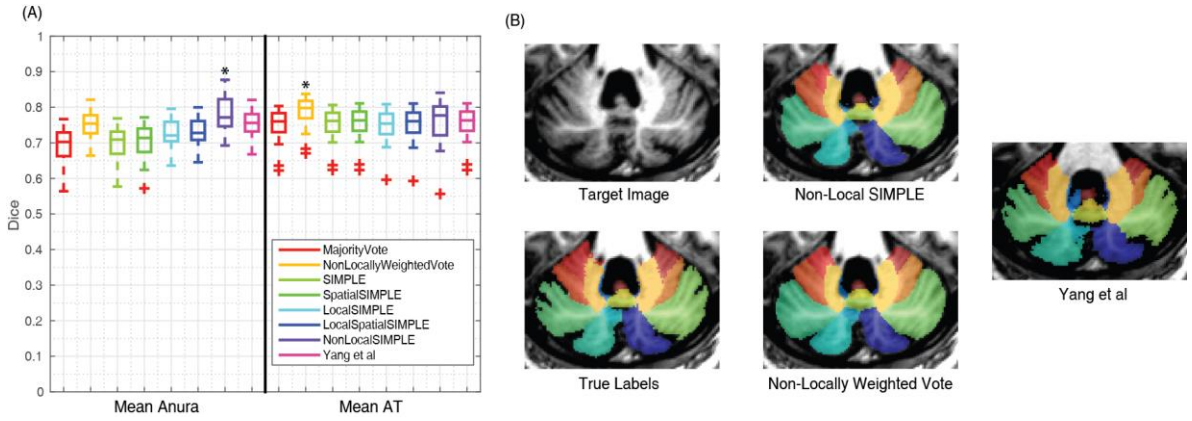

Figure V-2 Summarized segmentation results for the Anura and AT datasets. Non-Local SIMPLE outperformed all other techniques on the Anura dataset (A). On the AT dataset Non-Locally Weighted Vote significantly outperformed all other techniques, but Non-Local SIMPLE still outperformed the previously gold-standard technique of Yang et al (A).Qualitatively, Non-Locally Weighted Vote seemed to oversegment the lobules whereas Non-Local SIMPLE tended to undersegment. The results of Yang et al visually produced results more consistent with the anatomic boundaries but had more internal boundary shifts than either Non-Locally Weighted Vote or Non-Local SIMPLE.

30 patients with various cerebellar diseases, ranging in age from 29 to 90, 21 female 24 male, scanned with a 3T three dimensional MPRAGE sequence. Each subject was labeled by two intermediate experts and gold-standard segmentations were generated by fusing the manual labelings together.

2.2. Multi-Atlas Pipeline

All data from both populations followed the same protocol for registration. The data were first bias corrected with N4 bias correction. For each dataset, each pair of scans was non-rigidly registered using the Advanced Normalization Tools (ANTs) SyN algorithm and the default parameters for brain registration [88]. Labels volumes were then deformed to the subject space using the ANTs warping tool and nearest neighbor interpolation.

We compare our new algorithms with several previous algorithms. The first algorithms we compare against are majority vote and non-locally weighted [84, 96]. Second, we compare against

the SIMPLE algorithm from Langerak and a spatially varying extension, herein Spatial SIMPLE [105, 163, 164]. Lastly we compare our results to previous work on the same dataset by Yang et al where a multi-atlas segmentation was used as an initialization and a post-hoc graph cut was used to correct the image boundaries [165].

2.3. Local and Local Spatial SIMPLE

Following the generative model definition of SIMPLE from Xu et al [163] we incorporate local intensity into the model as

$$f(T_i = s, I_i | D_i, A_i, c, \epsilon, \sigma, \theta) \tag{2}$$

which we can solve through expectation-maximization. We define the E-Step as

$$\begin{aligned} W_{si}^k &= \frac{f(T_i = s) \prod_{j \in R} f\left(A_{ij} \middle| I_i, \sigma\right) f\left(D_{ij} \middle| T_i = s, c_j^k, \epsilon_j^k\right)}{\sum_{s''} f(T_i = s'') \prod_{j \in R} f\left(A_{ij} \middle| I_i, \sigma\right) f\left(D_{ij} \middle| T_i = s'', c_j^k, \epsilon_j^k\right)} \\ &= \frac{f(T_i = s) \prod_{j \in R} p\left(A_{ij} \middle| I_i, \sigma\right) \theta_{j c_j^k s s'}}{\sum_{s''} f(T_i = s'') \prod_{j \in R} f\left(A_{ij} \middle| I_i, \sigma\right) \theta_{j c_j^k s'' s'}} \end{aligned} \tag{3}$$

assuming conditional independence between the raters and the rater's intensity and where $p\left(A_{ij} \middle| I_i\right) = \exp \frac{\left(A_{ij} - I_i\right)^2}{\sigma}$ [84]. The M-Step directly follows Xu et al so it is excluded from this work. Briefly, the maximization of ϵ_j^{k+1} is total weight of the observed label for rater j across the image and c_j^{k+1} is defined based on a semi-parametric atlas selection method from the original SIMPLE definition [105]. To extend the model to be spatially varying we redefine θ as an $R \times 2 \times L \times L \times N$ matrix defined identically as before, c as an $N \times R$ matrix corresponding to

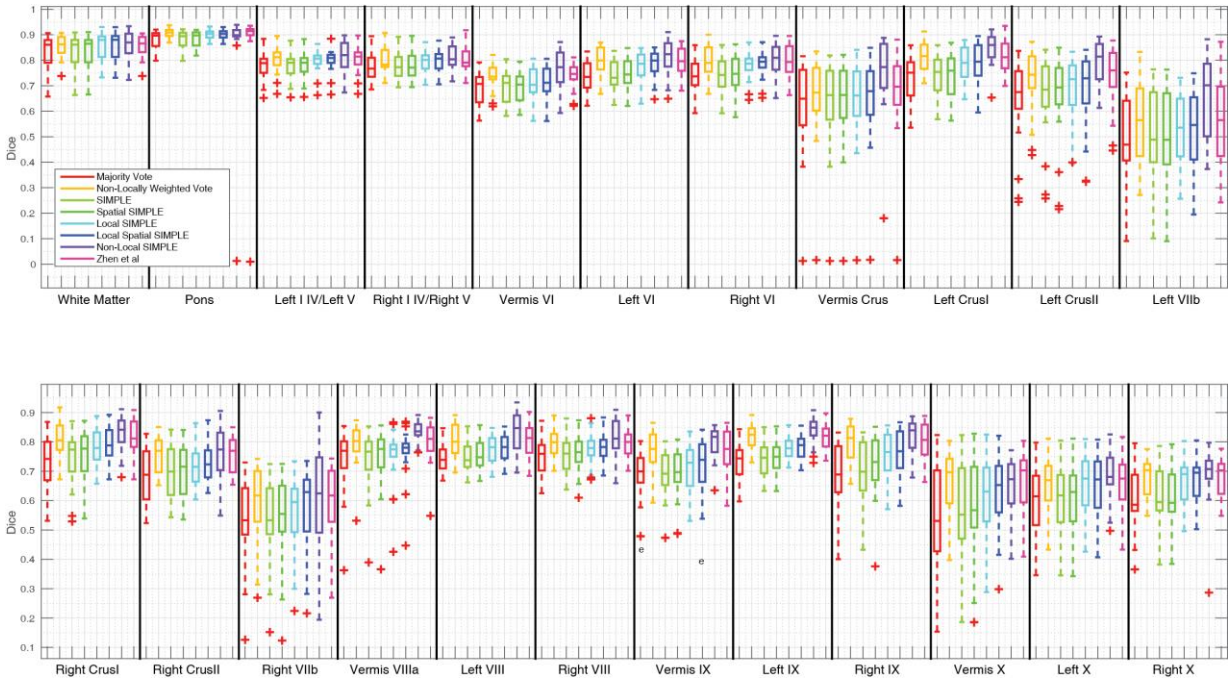

Figure V-3 Quantitative segmentation results for the Anura dataset. Non-Local SIMPLE shows either significant improvements over other algorithms or comparable results to other algorithms for all labels.

the atlas selection decision for each rater at each voxel, and ϵ as an $N \times R$ error vector for each rater at each voxel. The E-Step becomes

$$W_{si}^{k} = \frac{f(T_i = s)\prod_{j \in R} p\left(A_{ij} \middle| I_i, \sigma\right)\theta_{jc_{ji}^{k} s s' i}}{\sum_{s''} f(T_i = s'')\prod_{j \in R} f\left(A_{ij} \middle| I_i, \sigma\right)\theta_{jc_{ji}^{k} s'' s' i}} \tag{4}$$

and the M-Step once again follows Xu except the values of ϵ are calculated over the pooling are b and thus c is calculated per-voxel based on the estimates of ϵ.

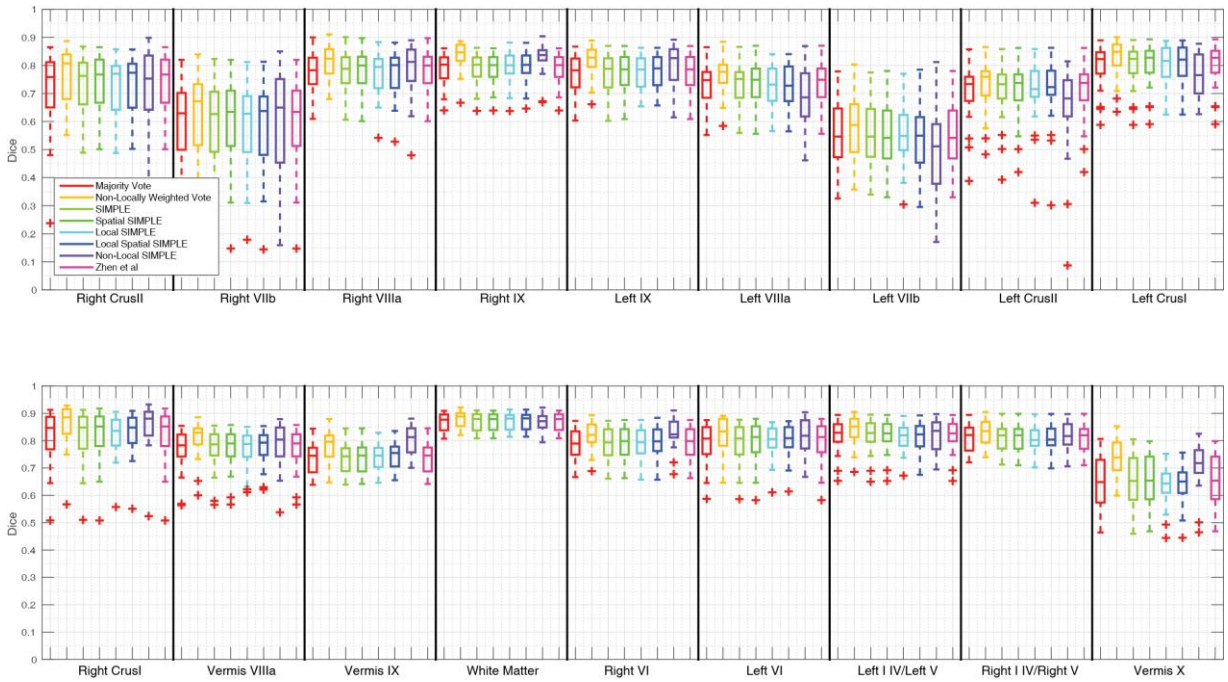

Figure V-4 Quantitative segmentation results for the Ataxia dataset. No algorithm shows significant improvement across all labels but Non-Locally Weighted Vote provides both consistent and accurate results across most labels.

2.4. Non-Local SIMPLE

Patch-based label fusion has been incorporated into many label fusion techniques such as Non-Local STAPLE and Non-Locally Weighted Vote [96, 97]. In these techniques, the correspondence model smooths the labels over the nearby region based on the intensity differences. We define the generative model of Non-Local SIMPLE as

$$f(T_i = s, I \mid D, A, c, \epsilon, \sigma, \aleph_s, \aleph_p, \theta, b) \tag{5}$$

where $\aleph_s$ are the parameters of non-local search, $\aleph_p$ are the parameters of non-local distance calculation, c as an $N \times R \times \aleph_S$ matrix corresponding to the patch selection decision for each rater at each voxel over their non-local search space, ϵ as an $N \times R \times \aleph_S$ error matrix for each rater at each voxel over their non-local search area, and θ is a confusion $R \times 2 \times L \times L \times N \times \aleph_s$ defined

both spatially and over the non-local correspondence search region. We estimate the solution of this model through expectation maximization. We define the E-Step as

$$W_{si}^{k} = \frac{f(T_i = s)\prod_{j\in R}\prod_{i'\in\aleph_S} p\left(A_{i'}\middle|I_i,\aleph_p,\sigma\right)\theta_{jc_{jii'}^{k}ss'ii'}}{\sum_{s''} f(T_i = s'')\prod_{j\in R}\prod_{i''\in\aleph_S} p\left(A_{i''}\middle|I_i,\aleph_p,\sigma\right)\theta_{jc_{jii''}^{k}s''s'ii''}} \tag{6}$$

where

$$p\left(A_{i'}\middle|I_i,\aleph_p,\sigma\right) = \exp\left(-\frac{\left|\left|\aleph_p(A_{i'}) - \aleph_p(I_i)\right|\right|^2}{\sigma}\right) \tag{7}$$

which is the standard definition of non-local correspondence of Euclidean distance between the atlas and target patches in an exponential distribution [96, 97]. This E-Step expansion assumes a conditional independence between patches and the non-local intensity probability model. The M-Step follows as with Local Spatial STAPLE where the confidence is calculated over the pooling region b between the patch in the atlas and the target voxel in the atlas. For instance

$$\epsilon_{iji'}^{k+1} = \arg_{\epsilon_{\mathrm{iji'}}}\max\sum_{s'}\sum_{a\in-b:b}\sum_{s} W_{s(i+a)}^{k}\ln\theta_{jc_{jii'}^{k}ss'ii'}\,\delta\left(D_{(i'+a)j},s'\right) \tag{8}$$

following from the derivation of Xu et al, where δ is the dirac delta function. Thus, Non-Local SIMPLE performs patch-based performance modeling with strong atlas selection following from the works of Langerak, Xu, and Coupe [105, 163].

2.5. Statistical Analysis

To assess the performance of each statistical fusion technique, each atlas was segmented in a leave-one-out study with each algorithm (i.e., 24 atlases per target in the Anura set and 44 atlases per target in the AT set). Since the registration and label propagation steps were identical

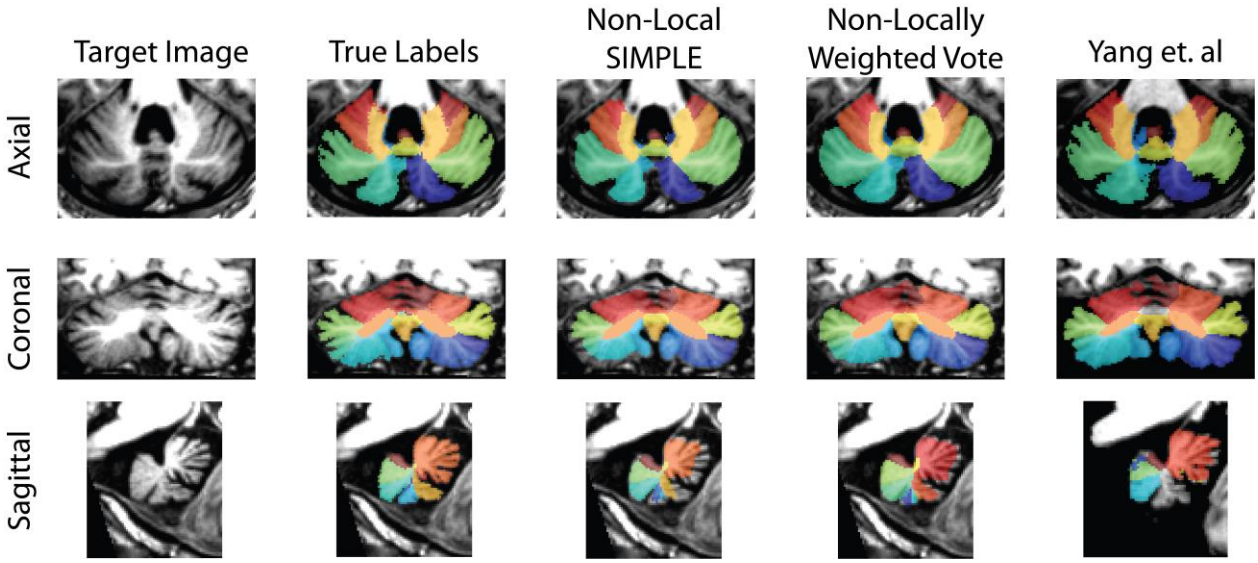

Figure V-5 Qualitative segmentation results from the median Ataxia subject. Non-Locally Weighted Vote tends to slightly over-segment regions of interest while Non-Local SIMPLE tends to under-segment regions. The adaptation of Yang et. al appears to generate a segmentation more consistent with anatomic boundaries but can produce severe missegmentations as seen in the sagittal view. Other algorithms are not shown since they infrequently outperformed the algorithms shown here.

between algorithms, we treat the segmentation results as paired between label fusion algorithms. We calculate the Dice coefficient between each set of true atlas labels and each label fusion approach. Since we cannot assume these Dice results fit any distribution, we perform a Wilcoxon signed-rank test between each algorithm. All significant results reported are at a $p<0.05$.

3. Results

In the leave-one-out segmentation of the Anura dataset, Non-Local SIMPLE produced statistically significant improvements in mean Dice compared to all other algorithms. Non-Local SIMPLE had an improvement in mean Dice of 0.03 on average compared with Non-Locally Weighted Vote and the approach of Yang et al. On the AT dataset, Non-Locally Weighted Vote significantly outperformed all other techniques by at least 0.04 mean Dice. Non-Local SIMPLE significantly outperformed the results of Yang et al on the AT data by 0.01 Dice (Figure 2).

Qualitatively Non-Locally Weighted Vote tends to slightly over-segment the cerebellar lobules whereas Non-Local SIMPLE under-segments. The results of Yang et al appear to produce segmentations more consistent with the true anatomic boundaries but have greater issues with labels shifting between regions (Figure 2). The full Dice scores for all regions of interest are available in Figures 3 and 4 and more qualitative results are available in Figure 5.

4. Discussion

In this work, we investigated new algorithms for fully-automated multi-atlas segmentation of the cerebellum. We proposed three approaches for segmentation deriving from the work of Langerak and Xu on the SIMPLE atlas selection and performance model [105, 163]. The first two algorithms, Local SIMPLE and Local Spatial SIMPLE, incorporated local image similarity into the generative model definition of SIMPLE and extended the base algorithm to consider only the local region in performance model calculation. The third algorithm, Non-Local SIMPLE, extends the SIMPLE model to patches in the area around the registered atlas images, incorporating the work of Coupe and patch-based segmentation into SIMPLE [96]. We then evaluated these algorithms against several previous algorithms, including the previous gold-standard cerebellar segmentation algorithm, on two sets of cerebellar atlases [165]. On the first set, Non-Local SIMPLE beat all other techniques with a $p < 0.05$. On the second set, Non-Locally Weighted Vote produced the best segmentation results, but Non-Local SIMPLE still outperformed the previous gold-standard technique. In conclusion, we have shown that cerebellar segmentation is a challenging task and no current technique produces significant improvements over other techniques so application specific considerations and trade-offs should be considered. Future work will investigate secondary processing techniques [99] to address systematic over/under-

segmentation concerns with the currently leading methods. We note that the proposed techniques are targeted at cases where a large number of atlases are available (i.e., greater than 30).

Chapter VI

Multi-Modal Imaging with Specialized Sequences Improves Accuracy of the Automated Sub-Cortical Grey Matter Segmentation

1. Introduction

The subcortical grey matter is a collection of nuclei situated near the forebrain [16]. These nuclei are primarily involved in connecting distinct portions of the brain, to serve as major functional systems within the brain [16]. For instance, the globus pallidus internal receives GABAergic signaling from the putamen and relays that to the sub-thalamic nucleus. Many subcortical structures have been implicated in one or more diseases [166]. In addiction, dopamine is dysregulated in the putamen and adjacent structures causing dependence phenotypes [167]. In Parkinson's disease, several subcortical structures undergo Lewy body growth and that growth plays a significant role in the motor phenotypes associated with the disease [168].

Recently, specialized imaging sequences have been developed for studying subcortical grey matter using clinical magnetic resonance (MR) scanners. In particular, the Fast Grey Matter Acquisition T1 Inversion Recovery (FGATIR) sequence was developed to improve subcortical grey matter contrast with the surrounding tissue [169]. The FGATIR sequence uses a longer inversion time than standard T1-weighted approaches, such as Magnetization Prepared Rapid Acquisition Gradient Echo (MPRAGE), to null the white matter and accentuate the deep brain structures. FGATIR images accentuate the sub-thalamic nucleus, lamina separating the internal and external globus pallidus, and the thalamus amongst other important structures. On the other

hand, many important subcortical structures are still difficult to parcellate using the FGATIR. Higher field strength scanners are needed, but these scans at higher field strengths are not clinically feasible in most contexts [114].

In this work, we investigate the efficacy of sequences acquirable in a clinically tolerable setting, namely standard T1-weighted MPRAGE scans and T1-weighted FGATIR scans acquired at 3T. The structures considered in this work are the substantia nigra (SN), subthalamic nucleus (STN), internal globus pallidus (GPI), external globus pallidus (GPE), putamen, and thalamus. These subcortical structures were manually delineated using a combination of scans acquired at 3T and 7T. Furthermore, we compare segmentation using only one modality, either the MPRAGE or FGATIR, to multi-modal segmentation using the enhanced and complimentary contrast patterns present to improve the overall segmentation results.

2. Methods

We propose a multi-atlas segmentation algorithm for automated segmentation of the subcortical grey matter. This approach uses multi-modal atlases derived using imaging acquired at 3T and 7T. The segmentation uses the imaging sequences acquired at 3T to assess the effectiveness of segmenting subcortical structures using clinically feasible acquisitions.

2.1. Atlas Imaging

Nine healthy subjects were scanned at 3T and 7T. At 7T, a series of 0.7mm isotropic T1-weighted MPRAGE (Inversion Time (TI)/ Repetition Time (TR)/Echo Time (TE) =[400,640,960,1120]/4.74/2.09ms) was acquired and a susceptibility weighted image slab through the midbrain acquired at 0.2x0.2x1.1mm was acquired sagitally, coronally, and axially (TR/TE/Flip Angle (FA)=1952/23ms/45° for all orientations). At 3T, a 1.0mm isotropic resolution

T1-weighted MP-RAGE (TI/TR/TE=925/8.1/2.7ms) and an FGATIR scan was acquired for additional mid-brain contrast (TI/TR/TE=400/7.39/3.43).

2.2. Manual Segmentation

For each subject, the 7T T1-weighted MP-RAGE with the inversion time of 960ms was used as the reference space. The other 7T MP-RAGE scans, 3T MP-RAGE, 7T high-resolution susceptibility weighted slabs, and 3T FGATIR were co-registered to the reference space. The following structures were manually labeled on the left hemisphere for one subject: GPI, GPE, STN, SN, thalamus, and the putamen. The labeled atlas was then registered to each of the other eight subjects and the labels were deformed to the target space using the Reg Aladin algorithm in NiftyReg [170]. The deformed labels were then manually corrected. Finally, each subject was flipped laterally and the flipped image was registered to the standard space image. Each subject's labels were deformed in the laterally flipped space and the results were manually corrected. The final result was nine subjects with left and right labels for the GPI, GPE, STN, SN, thalamus, and putamen. All manual segmentations were done using CranialVault and the CRAVE Tools [171].

2.3. Segmentation Algorithm

First, each subject, the 3T T1-weighted MRI was automatically segmented with the BrainCOLOR (www.neuromorphometrics.com) following a standard multi-atlas segmentation approach [89]. Briefly, the target image was affinely registerd to MNI space. From a population of 45 atlases, the 15 atlases geodesically most similar to the target are then selected. These 15 atlases are non-rigidly registered to the target image using the Advance Normalization Tools (ANTs) Symmetric Normalization algorithm (SyN) [87]. Finally, the registered atlas images and labels are fused to the target image using Hierarchical Non-Local Spatial STAPLE [93].

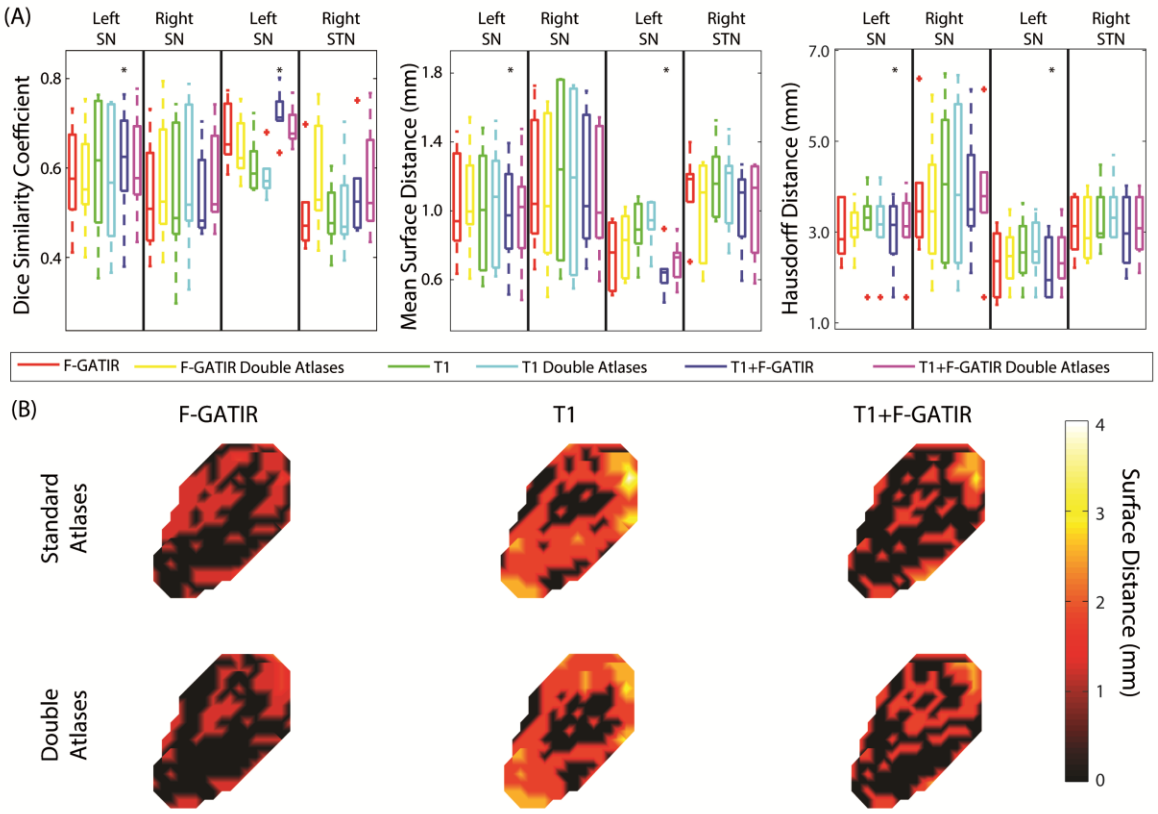

Figure VI-1 Segmentation results for structures in the dienchephalon. Quantitative segmentation results are shown in (A). For the left SN, multi-modal segmentation with T1 and FGATIR outperformed other approaches (*; $p<0.05$; Wilcoxon sign-rank test). For the right SN no segmentation approach outperformed other approaches. For the left STN, multi-modal segmentation with T1 and FGATIR outperformed other approaches (*; $p<0.05$; Wilcoxon sign-rank test). For the right STN no segmentation approach outperformed other approaches. In (B), surface distances between the true and estimated segmentations for the left SN are shown for the six proposed segmentation approaches.

Whole brain segmentation (WBS) was used to localize the particular regions of interest. In particular, the thalamus label from the WBS was used to localize the thalamus, the globus pallidus and putamen labels from the WBS were used to localize the GPI, GPE, and putamen, and the diencephalon label from the WBS was used to localize the SN and STN. The bounding box of each of these regions of interest was identified and dilated by 5mm. Finally, the labels and T1 and

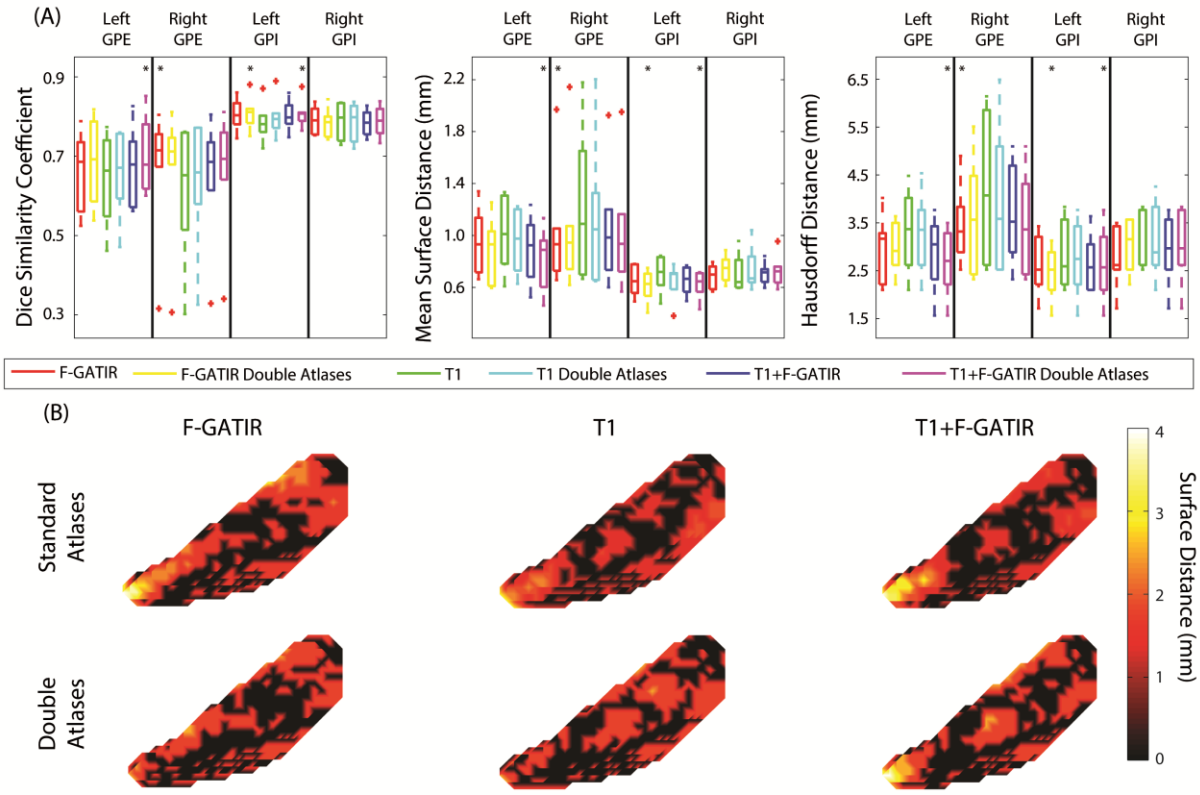

Figure VI-2 Segmentation results for structures in the globus pallidus. Quantitative segmentation results are shown in (A). For the left GPE, multi-modal segmentation with T1 and FGATIR with double atlases outperformed other approaches (*; $p<0.05$; Wilcoxon sign-rank test). For the right GPE segmentation with FGATIR outperformed other approaches (*; $p<0.05$; Wilcoxon sign-rank test). For the left GPI, multi-modal segmentation with T1 and FGATIR with doubled atlases and segmentation with FGATIR with doubled atlases outperformed other approaches but were not distinguishable amongst each other (*; $p<0.05$; Wilcoxon sign-rank test). For the right GPI no segmentation approach outperformed other approaches. In (B), surface distances between the true and estimated segmentations for the left GPI are shown for the six proposed segmentation approaches.

FGATIR intensities were extracted from these bounding boxes and saved as reduced field of view (RFOV) atlases.

For a given target, the target was segmented with the BrainCOLOR protocol (www.neuromorphometrics.com). A series of targets (RFOV) were created following the protocol defined above. The RFOV atlases were co-registered to the RFOV targets. All registrations were performed using with ANTs and the SyN algorithm [88]. After registration, joint label fusion (JLF)

was used. In all cases, the same collection of imaging modalities was used for the segmentation [98]. Finally, each structure's segmentation was reinserted into the standard image space. All operations for creating and manipulating using RFOV atlases and images used custom MATLAB (www.mathworks.org) code.

3. Results

Each of the nine healthy subjects was segmented in a leave-one-out cross validation scheme. First, each subject was segmented using the T1-weighted MRI, the FGATIR, and multi-modally with the T1 and FGATIR. Second, each subject's scans were flipped left-right to produce a second set of atlases. Each subject was then segmented with the 16 atlases, leaving out the atlas and the flipped version of the atlas. As a result, each subject was segmented six times, twice with each combination of modalities. The results are divided into three pieces for ease of visualization: diencephalon (STN and SN), GPI and GPE, and thalamus and putamen. For each segmentation result the Dice Similarity Coefficient (DSC), mean surface distance (MSD), and Hausdorff distance (HD) were calculated.

3.1. Diencephalon

Four structures were segmented in the diencephalon: the left STN, right STN, left SN, and right SN (figure 1). For the left SN, the segmentation with T1 and FGATIR outperformed other approaches ($p<0.05$ Wilcoxon sign-rank test) with a median DSC of 0.65, median MSD of 0.98 mm, and a median HD of 3.11 mm. For the right SN, no approach significantly outperformed other approaches. For the left STN, segmentation with T1 and FGATIR outperformed other approaches ($p<0.05$ Wilcoxon sign-rank test) with a median DSC of 0.70, median MSD of 0.61 mm, and a

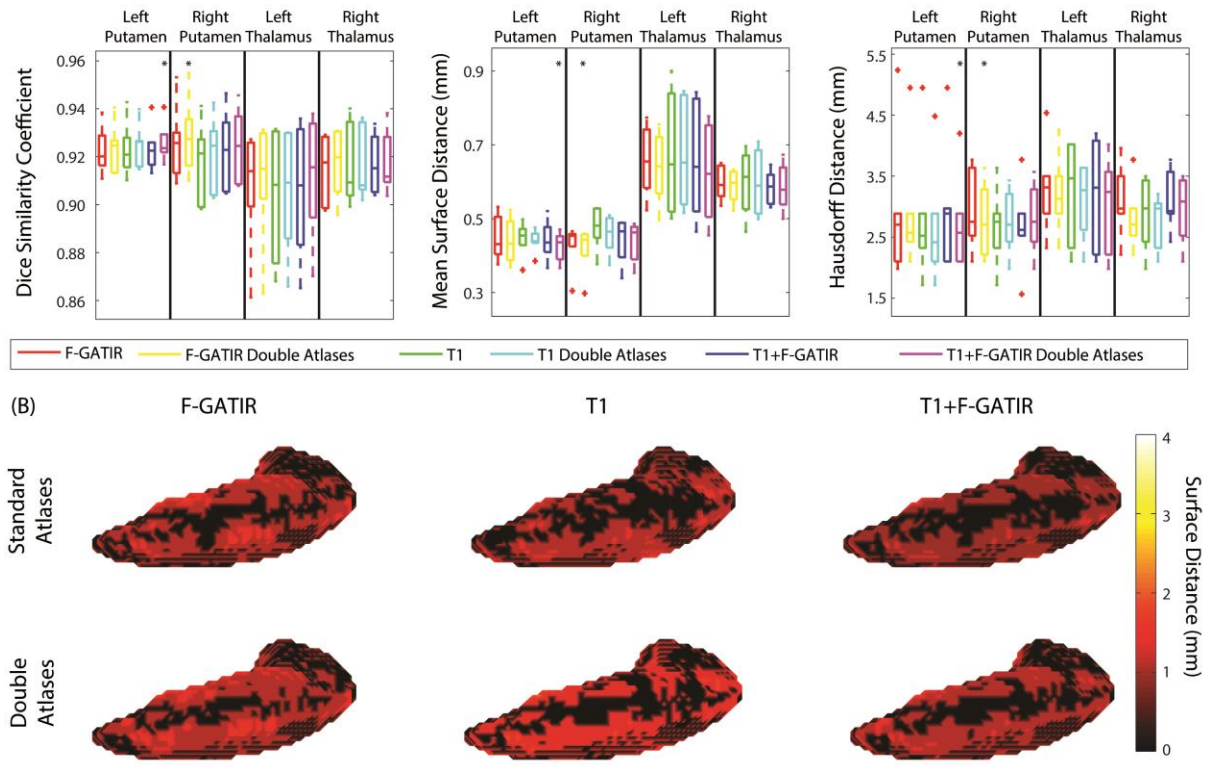

Figure VI-3 Segmentation results for the putamen and thalamus. Quantitative segmentation results are shown in (A). For the left putamen, multi-modal segmentation with T1 and FGATIR with double atlases outperformed other approaches (*; $p<0.05$; Wilcoxon sign-rank test). For the right putamen segmentation with FGATIR with doubled atlases outperformed other approaches (*; $p<0.05$; Wilcoxon sign-rank test). For the left thalamus, no segmentation approach outperformed other approaches For the right thalamus, no segmentation approach outperformed other approaches. In (B), surface distances between the true and estimated segmentations for the left putamen are shown for the six proposed segmentation approaches.

median HD of 2.06 mm. For the right STN, no approach outperformed other approaches (Wilcoxon sign-rank test).

3.2. Globus Pallidus

Two structures were segmented in the globus pallidus: the GPI and GPE. These structures were segmented bilaterally and resulted in four total structures segmented (figure 2). For the left GPE, the segmentation with T1 and FGATIR including flipped atlases outperformed other

approaches ($p<0.05$ Wilcoxon sign-rank test) with a median DSC of 0.68, median MSD of 0.94 mm, and a median HD of 2.70 mm. For the right GPE, segmentation with FGATIR outperformed other approaches ($p<0.05$ Wilcoxon sign-rank test) with a median DSC of 0.71, a median MSD of 0.96 mm, and a median HD of 3.42 mm. For the left GPI, segmentation with FGATIR with flipped atlases and multi-modal segmentation with T1 and FGATIR and flipped atlases outperformed other approaches but were not statistically distinguishable from each other ($p<0.05$ Wilcoxon sign-rank test), with median DSC values of 0.80 and 0.81, median MSD values of 0.68 and 0.69 mm, and median HD values of 2.50 and 2.52 mm respectively. For the right GPI, no approach significantly outperformed another.

3.3. Thalamus and Putamen

The left and right thalamus and putamen were segmented resulting in four total structures (figure 3). For the left putamen the segmentation with T1 and FGATIR including flipped atlases outperformed other approaches ($p<0.05$ Wilcoxon sign-rank test) with a median DSC of 0.93, a median MSD of 0.42 mm, and a median HD of 2.50 mm. For the right putamen, segmentation with FGATIR including flipped atlases outperformed other approaches ($p<0.05$ Wilcoxon sign-rank test) with a median DSC of 0.93, a median MSD of 0.45 mm, and a median HD of 2.61 mm. For the left thalamus, no approach significantly outperformed the others. Finally, for the right thalamus, no approach significantly outperformed the others.

4. Discussion

In this work, we presented segmentation approaches for segmenting six subcortical structures bilaterally. These segmentation approaches considered the effect of imaging modality on segmentation results. Two distinct imaging modalities were considered. First, a standard T1-

weighted MPRAGE, a sequence commonly acquired in clinical and research settings, was acquired for nine subjects. Second, a T1-weighted FGATIR, a specialized sequence with enhanced contrast in subcortical structures, was acquired for the same nine subjects. A series of 7T T1-weighted MPRAGE scans with varying inversion times and high-resolution susceptibility weighted slabs were acquired on the nine subjects. Then, an expert in subcortical anatomy manually delineated the thalamus, putamen, internal and external globus pallidus, sub-thalamic nucleus, and substantia nigra bilaterally.

These nine subjects were then used in a leave-one-out cross-validation to assess the segmentation accuracy using only T1-weighted MPRAGE, only T1-weighted FGATIR, and multi-modally with the MPRAGE and FGATIR. In general the multi-modal segmentation outperformed the other approaches and furthermore including atlases flipped laterally tended to improve segmentation results, but there was no single approach that outperformed in all cases. Fortunately, the proposed segmentation approach does not require all segmentations to be performed with the same modalities. This allows flexibility in which sequences are used to segment each structure.

These sequences are of interest because they are all acquirable on a clinical population. As a result, the proposed segmentation approaches can be translated to clinical populations and thus aid in the clinical workflow. In particular, the STN and GPI are common targets for deep brain stimulation surgery (DBS) [67]. DBS is a surgery commonly used in Parkinson's disease to mitigate the motor symptoms of the disease. Overall, this work is a significant step in understanding the effects of imaging sequence on segmentation of subcortical grey matter structures.

Chapter VII

Improving Variance Estimation in Multi-Atlas Segmentation to Accurately Characterize the Marginal Utility of Additional Atlases

1. Introduction

One of the most common goals of medical image processing is automated segmentation of regions of interest (ROIs) on structural imaging [89]. The goal of automated segmentation is to delineate the location in the image containing the ROI for a given target image [89]. For instance, given a magnetic resonance image (MRI) of the human brain, a common ROI of interest is the hippocampus [172]. Automated segmentation of the hippocampus seeks to accurately and reproducibly delineate the hippocampus from the surrounding tissue and background.

One of the most popular and translatable techniques for automated segmentation is multi-atlas segmentation [89]. In multi-atlas segmentation, a set of atlases (i.e., representative images with the structure or structures of interest delineated) is registered to a target image to be segmented. These registered atlases are then joined together to create a consistent representation of the target structure. The step of joining the atlases together is commonly referred to as "fusion" or "label fusion." The goal of multi-atlas segmentation is to produce a consensus representation, ideally more consistent with the truth than any individual atlas could produce. The multi-atlas framework has been applied to structures ranging from the brain [173], to the optic nerve [174], to the abdomen [150], and extendible outside of humans [89].

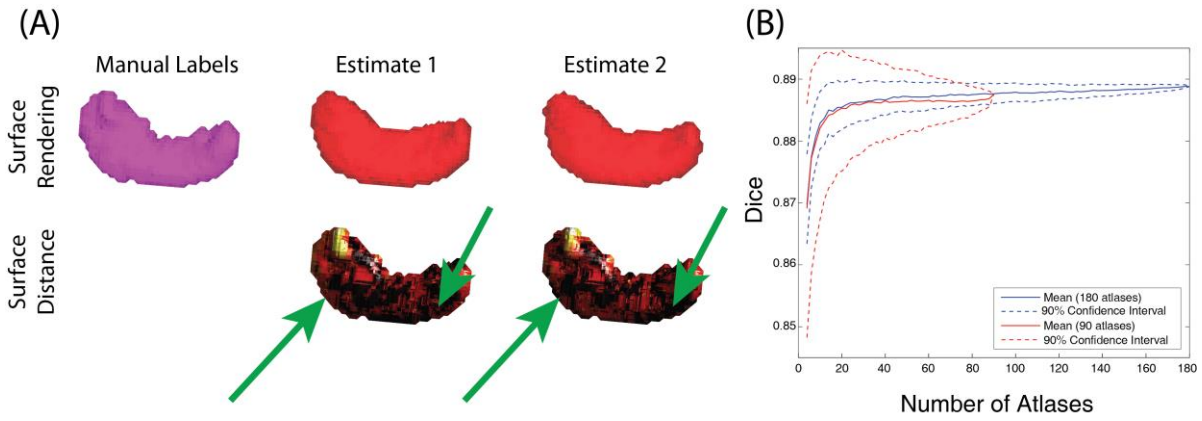

Figure VII-1 Example segmentations of the hippocampus using 20 atlases (A). Estimate 1 and Estimate 2 use unique populations of atlases. The surface estimates show significantly different estimates at several boundary locations (identified by the green arrows). Monte Carlo estimates of confidence intervals become invalid especially as the number of atlases in the sample approaches the number of atlases available (B). In cases where the total population size is 100 (red curves) and 200 (blue curves), the variance estimated by Monte Carlo approached 0 as the number of atlases used approached the total pool size.

The availability of atlases covering the anatomical diversity in the target population is important to accurately perform the segmentation. Due to the computational complexity of the registration [175] and fusion steps, determining the minimum number of atlases to accurately segment a target image is important for the development of multi-atlas algorithms.

Several metrics are used to estimate the accuracy of multi-atlas algorithms and are commonly used to compare various segmentation approaches. Each of these approaches uses manually defined labels as a baseline for assessing the accuracy. The first approach, Dice Similarity Coefficient (DSC), measures the voxel-wise overlap between the true and estimated segmentations [102]. In particular, it measures the magnitude of the intersection of the truth and estimate divided by the size of the true and estimated segmentation. DSC has come against some criticism since large ROIs, such as the white matter in the brain or the liver in the abdomen, will have artificially high DSC values compared to that of smaller structures. Two separate approaches for estimating segmentation accuracy utilize surfaces of the labels. The first approach, mean

surface distance (MSD) calculates the average distance between the true and estimated segmentation. Hausdorff distance (HD) is the maximum distance between the true and estimated surface [176]. These two metrics together give a stronger representation of the accuracy of the segmentation since they are not biased by the size of the structure.

Typically, analyses of multi-atlas segmentation algorithms perform a leave-one-out cross validation to validate a particular approach against other techniques. In these leave-one-out approaches, N-1 atlases, where N is the total number of atlases available, are used to segment the remaining one atlas. DSC, MSD, and HD values are then calculated against the truth. This process is repeated for each atlas and for each segmentation approach. Each of the segmentation approaches is then compared with a paired t-test or a Wilcoxon sign-rank test to evaluate if there is a significant difference between approaches.

In order to determine the number of atlases needed for a segmentation task, previous works have selected a random subset, k, of the atlases available and performed a segmentation of the target with the selected atlases. Then, accuracy measures are calculated on the data and a t-test or sign-rank test is used to determine the point where adding additional atlases does not improve the marginal segmentation results (e.g., [177]). This can be seen as a special case of the above framework where each approach considered is a particular number of atlases.

This approach is limited in that it does not consider the variability in segmentation results with respect to the population of atlases selected. Two segmentations of the same target with distinct populations of atlases and the same DSC, MSD, and HD can exhibit different patterns of error that are not captured in the summary measures (Figure 1A). Furthermore, the approach from [177] considers the population of target images but does not model the variance of an individual

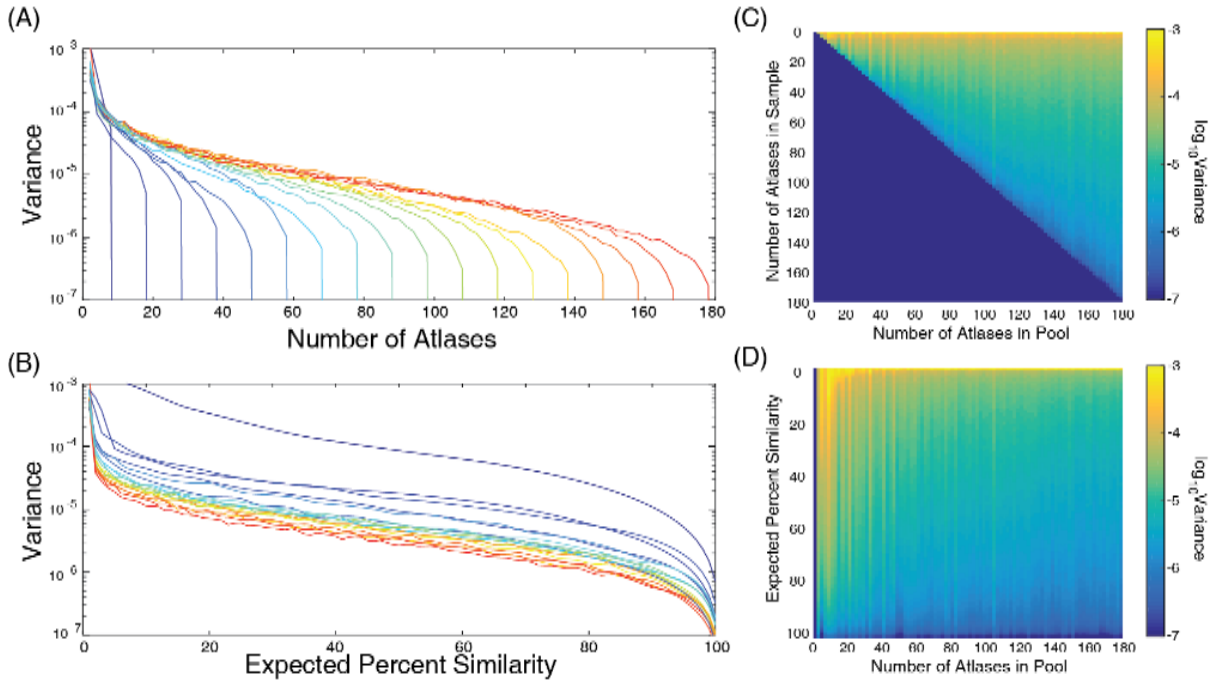

Figure VII-2 Variance estimation results from Monte Carlo sampling. In (A) each curve represents a pool of atlases of increasing size. The number of atlases are sampled from the pool in increasing amounts. These results are then interpolated with Expected Percent Similarity (B) to identify a consistent pattern of decline of variance with respect the Expected Percent Similarity. (C) and (D) show the patterns of (A) and (B) with all sizes of pools sampled.

subject. By not considering each evaluation subject as an individual, summary statistics may not capture that a particular subject has not converged in the population.

In this work, we present an algorithm for estimation of an individual subject's DSC with respect to the number of atlases used. We first show that using standard Monte Carlo sampling of atlases from a population does not properly capture the variance of the segmentation with respect to the number of atlases used (Figure 1B). We propose a mathematical solution to this using increasingly large populations of atlases that we draw our pool from. We then translate this to voxel-level results where we estimate the likelihood of a given label at each voxel. Following the same formulation for DSC, we estimate the variance of each voxel's label decision. From this result, we estimate the number of voxels likely to change given two distinct populations of atlases.

Given this result, we identify the number of atlases which minimizes the number of voxels likely to change between two segmentations as the number of atlases needed for the segmentation task.

2. Theory

In this section, we outline the theory underlying variance estimation. This theory is applicable to both estimation of variance when measuring DSC and measuring the distribution of a voxel's label decision. This section first outlines Monte Carlo sampling approaches and describes the variance estimation approach using Monte Carlo sampling. Finally, this section describes two alternate approaches for variance estimation that are only valid within certain restrictions. In the scope of this section, an estimator is considered to be a single registered atlas.

2.1. Monte-Carlo Sampling

In a Monte Carlo sampling, we begin with a pool of estimators, P, which consist of all of the available estimators for a particular task. Then, a sample of size n estimators is drawn without replacement from P. This procedure is repeated a number of times and can be used to estimate first-order statistics about the distribution of interest. Estimates of second-order and higher statistics are biased by this sampling procedure since the samples of size n drawn from P are correlated, and thus the variance estimated is lower than the true variance (Figure 1B). this procedure can be repeated for values of n from 2 to P to estimate accuracy and variance with n estimators for a population size of P.

2.2. Variance Estimation through Monte-Carlo Sampling

The algorithm to estimate the true variance with respect to the number of estimators proceeds as follows. First, a pool of size p where $p \leq P$ is drawn randomly from the total population P and without replacement from the population of estimators. Then, the procedure

outlined in *A* can be performed to estimate the performance of the estimators with pool size p. The result of this procedure is mean and variance estimate for counts of estimators and pool size ranging from 2 to P. Specifically, for a pool size of 100 estimators, this procedure results in mean and variance estimates for estimator counts ranging from 2 to 100.

Next, we introduce expected percent similarity (EPS) which is the percent of the estimators in common of a sample of size n from a population of p estimators. This measure is defined as

$$EPS_{np} = \frac{1}{p}\sum_{i=0}^{n}\frac{\binom{n}{i}\binom{p-i}{i}}{\binom{p}{s}} \quad (1)$$

where EPS_{np} is the EPS of n samples from a pool of size p. The intuition behind this that for every size up to the total number of samples, determine the likelihood that there are that many samples in common.

Thus, for a given sample size, irrespective of pool size, several estimates of the variance have been estimated with differing percent similarities. We propose to fit this function with a sigmoid function, which is a natural fit for this application since, when the number of estimators in the sample is near the total pool size, the variance will be near zero, but when the number of estimators in the sample is dramatically smaller than the pool size, then the variance will asymptotically approach the true variance with an infinite pool. This function is modeled as

$$EPS_{pn} = P_1 + \frac{P_2 - P_1}{1 + P_3 V_{pn}^{P_4}} \quad (2)$$

where P_{1-4} are parameters estimated in the regression, V_{pn} is the estimated variance of n estimators from the Monte Carlo procedure for a pool of size p, and the equation is fit for a fixed value of n. Though this function is solvable for EPS_{pn}, for the purpose of fitting the model this form is more intuitive for parameter initialization. This function is then solved for V_{pn}, and the

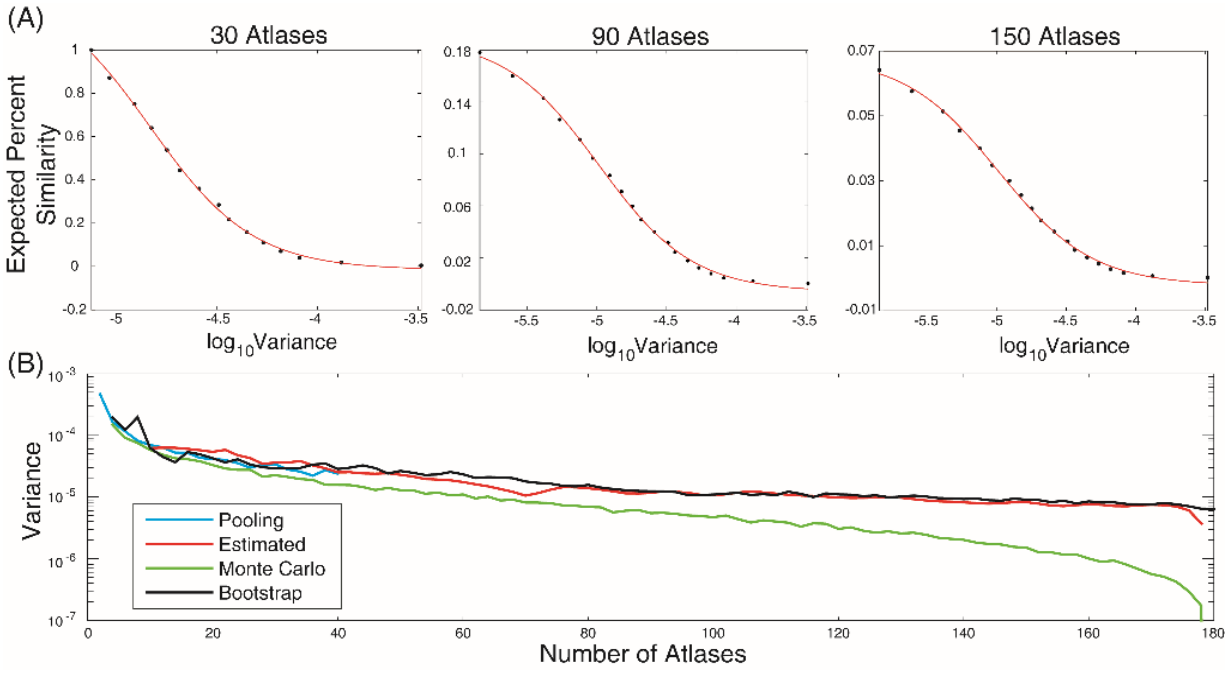

Figure VII-3 Example fits of the Expected Percent Similarity to the variance estimate (A) for increasing numbers of atlases. Variance values are scaled to their log for visualization purposes. (B) the final fit results for the estimated variance from the proposed algorithm, compared with the pooling approach, standard Monte Carlo, and bootstrapping approaches. The estimated approach follows a similar trend to the bootstrapping and pooling based estimates. On the other hand, the Monte Carlo approach deviated from other approaches in particular as the number of atlases considered approached the total available.

function is evaluated at a value near 0 EPS to estimate the true variance when there is little or no similarity between the atlases. A value near zero is used because the solution of (2) is possibly non-real for EPS values of 0, so a value near 0 is used to maintain numerical stability. This process is repeated for every sample size n to determine the true variance with n atlases.

2.3. Bootstrap

One of the common techniques to estimate a generic function given a limited pool of estimators is the bootstrap[178]. In bootstrapping, a sample of atlases of size n is drawn with replacement from the pool of estimators available. The bootstrap estimate of variance with a particular number of estimators is then calculated based on the results. This procedure is unbiased

unlike the Monte Carlo approach because the expected similarity of any two samples is zero. Bootstrapping is valid only when the estimator does not assume independence. For an approach like majority vote, this is a valid assumption, but for many modern segmentation approaches such as joint label fusion (JLF), independence needs to be assumed between atlases used. As a result, bootstrapping cannot generally be used for variance estimation with modern segmentation approaches.

2.4. Pooling

A separate technique for parameter estimation, referred to herein as pooling, uses a technique similar to bootstrapping, but does not sample with replacement. In estimation of variance with n atlases with a pool size of N, $\text{floor}\left(\frac{n}{N}\right)$ samples are drawn, where each sample is independent of the other samples. Variance is then estimated across the $\text{floor}\left(\frac{n}{N}\right)$ samples. This process is then repeated and averaged to produce an estimate more accurate than any one iteration [179]. This approach has shown to produce accurate first-order approximations of features. Since this approach estimates the variance as the mean of a series of estimations, it is a valid first-order estimate of variance. This process is applicable to any segmentation algorithm, but is limited in estimating the variance since a sufficient number of data points are needed to estimate variance.

3. Methods

For this work, 190 atlases labeled with the right hippocampus were considered. All atlas subjects are healthy controls and were sequenced with a 3D T1-weighted MPRAGE (TI/TR/TE = 860/8.0/3.7 ms; 170 sagittal slices; voxel size = 1.0mm3). Each scan was labeled following the protocol defined in [152]. Two studies were considered using this data. First, one subject was segmented using majority vote (MV) and the variance of the DSC of the segmentation with respect

to the truth and number of atlases used was examined. Second, ten atlases were randomly selected as a validation population. These ten atlases were segmented using JLF and voxel-wise variance estimates were established. These experiments are described in detail below.

3.1. Dice Similarity Coefficient Variance Estimation with Majority Vote

From the 190 atlases, one subject was randomly selected as the target subject. Of the remaining 189 atlases, 180 were selected as an atlas pool. The atlases were then registered to the target image affinely with the NiftyReg algorithm [86] and non-rigidly with the Symmetric Normalization (SyN) algorithm in the Advanced Normalization Tools (ANTs) package [88].

After registration was completed, the procedure described in §II.B was performed on the dataset. The procedure was performed with pool sizes of two to 180 in steps of two and sample sizes of two up to the pool size. For each step and pool size, the Monte Carlo sampling procedure was performed 1000 times to estimate the distribution (Figure 2). At each iteration, a sample of atlases was randomly drawn from the pool of atlases and MV was used to estimate the target. For each resulting segmentation, the DSC between the estimate and the target segmentation was calculated. Then the variance with respect to the number of atlases and the pool size was calculated. Once the Monte Carlo variance estimation was completed, Eq. 2 was fit based on the variance estimates and the variance with zero similarity was calculated (Figure 3A). The fitting was not performed for sample sizes less than 10 because there were not enough data points available to fit the four parameters.

Since majority vote does not require independence of the atlases, the bootstrapping procedure described in §II.C was performed. The full population of 180 atlases was used as a pool and sample sizes of two up to 180 in steps of two were used in the bootstrapping. For each sample size, 1000 bootstrap repetitions were performed to estimate the variance. The pooling procedure

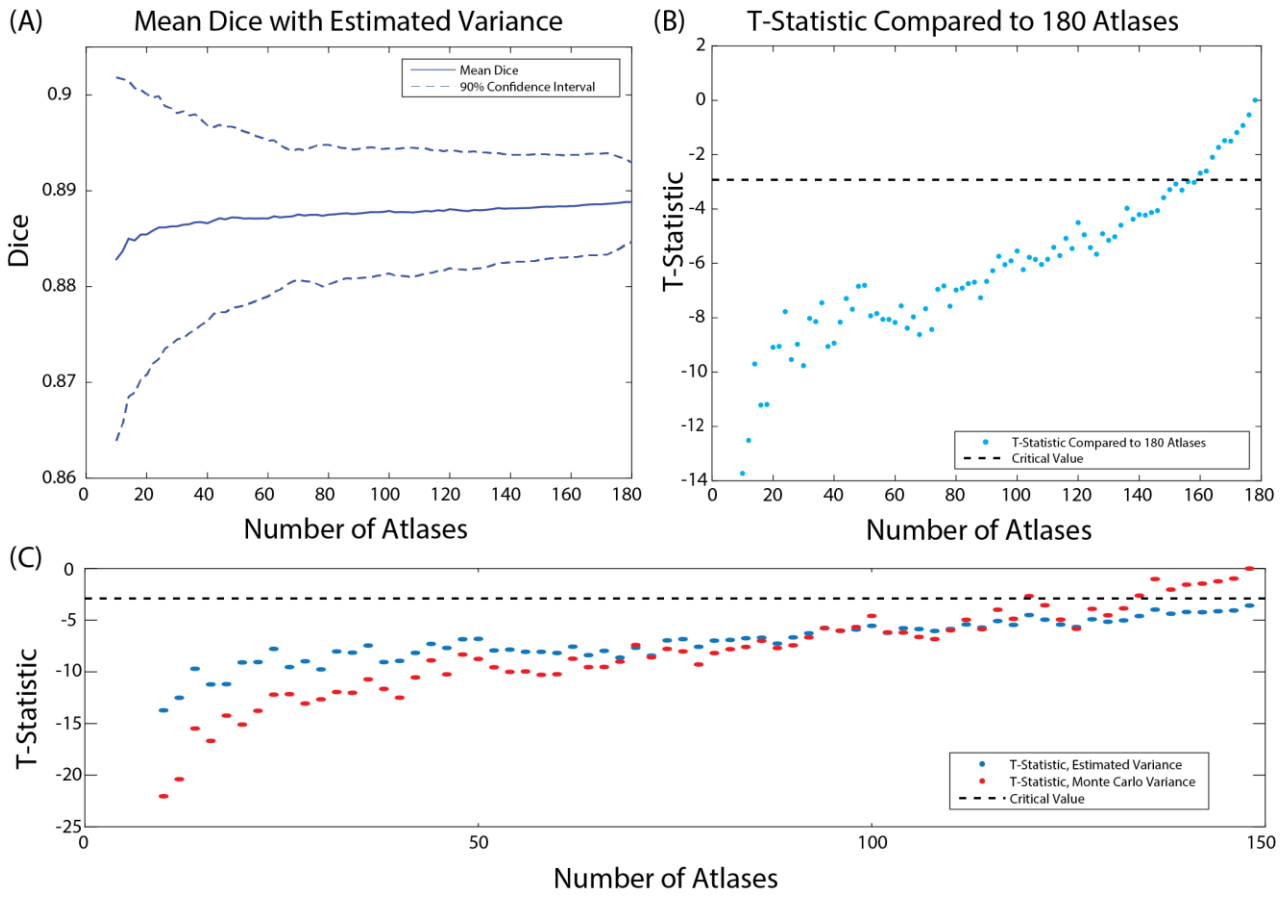

Figure VII-4 Re-fit variance results for the Monte Carlo estimation approach (A). These variance estimations do not converge as the standard Monte Carlo variance estimation did. These results show that variance is still decreasing as the number of atlases increases, and the average is also increasing. Comparing the final distribution with 180 atlases to the distributions with fewer atlases, we show that 180 atlases outperforms atlas counts up to 160 (B). Furthermore, when comparing the results for identifying the proper number of atlases from the population size of 150, the estimated variance approach identified that at least 150 atlases were needed whereas the Monte-Carlo variance approach identified 136 as the optimal number of atlases (C).

described in §II.D was also performed using the full population of 180 atlases. The pooling procedure was performed to a sample size of 40 atlases since multiple data points are needed to estimate the variance properly.

3.2. Voxel-Wise Variance Estimation with Joint Label Fusion

From the 190 atlases, 10 were selected as a target population. Of the remaining 180 atlases, 100 were randomly selected as a pool. The 100 atlases were then registered to the target images affinely with the NiftyReg algorithm [86] and non-rigidly with the Symmetric Normalization (SyN) algorithm in the Advanced Normalization Tools (ANTs) package [88].

After registration was completed, the procedure described in §II.B was performed on the datasets. The procedure was performed with pool sizes of two to 100 in steps of two and sample sizes of two up to the pool size. For each step and pool size, the Monte Carlo sampling procedure was performed 1000 times to estimate the distribution (Figure 2). At each iteration, a sample of atlases was randomly drawn from the pool of atlases and JLF was used to estimate the target.

For each resulting segmentation, the label-wise probability volumes were calculated and used to establish label-wise variance maps with respect to the number of atlases used and the pool size. The voxel-wise variance with respect to the number of atlases used was then calculated following §II.B and by fitting Eq. (2) for each voxel. Finally, these voxel-wise results were used to determine the number of voxels likely to change labels between two random draws of atlases from an infinitely large pool.

The bootstrap was not considered on this experiment because JLF requires atlases to be conditionally independent, which is violated in the boostrap. Also, this experiment was not extended beyond 100 atlases due to computational expense of the experiment. In particular, this experiment took over 400 CPU months to execute.

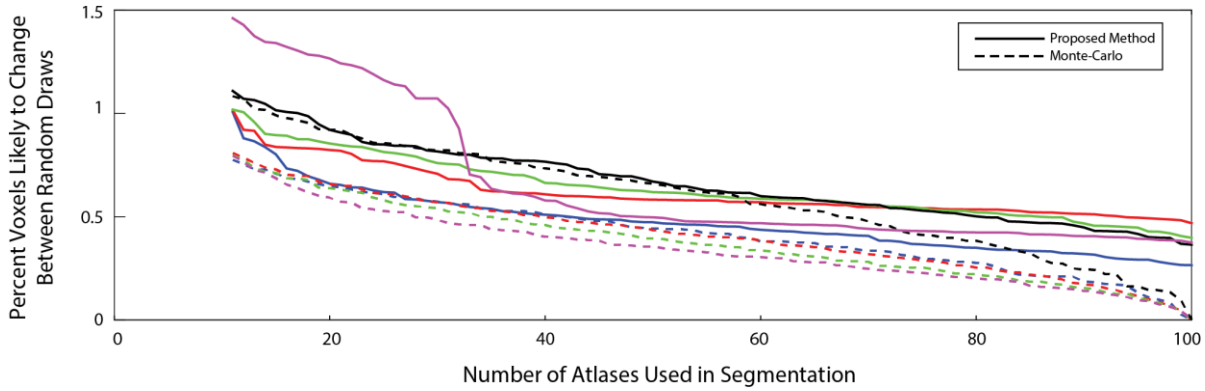

Figure VII-5 Percent of the total voxels likely to change between two unique draws of atlases from a population. The x-axis shows increasing numbers of atlases in the segmentation. Each color represents one of the subjects on which the experiment was performed. The filled in lines represent the estimates from the proposed approach, whereas the dotted lines represent the Monte Carlo estimates. We can see that as more atlases are added, a smaller portion of voxels are likely to change between draws from the atlas population. The Monte Carlo estimates tend to underestimate the number of voxels likely to change. The proposed method shows that, though the rate of change is decreasing of the number of voxels likely to change between two random draws, that there is still added value by adding additional atlases.

4. Results

4.1. Dice Similarity Coefficient Variance Estimation with Majority Vote

Four procedures for variance estimation were performed: standard Monte Carlo sampling (§II.A), pooling (§II.D), bootstrapping (§I.C), and our proposed variance estimation technique (§II.B). Since the atlas fusion technique proposed does not require independence of samples, the bootstrap estimate of variance is a valid estimation of variance. The approximation of variance estimate from the pooling is also not biased by the repetition of atlases. The bootstrap estimate of variance and the proposed estimated variance were significantly correlated (R^2=0.87, p<0.01). The bootstrap estimate of variance and the pooling estimate of variance were significantly correlated (R^2=0.79, p<0.01). The pooling estimate of variance and the proposed estimate of variance were significantly correlated (R^2=0.77, p<0.01). The Monte Carlo estimate of variance was not

significantly correlated with the bootstrap estimated of variance (R^2=0.40, p>0.1) or the proposed estimate of variance (R^2=0.46, p>0.1).

After the variance estimates were established, we used the Monte Carlo estimates of mean Dice and the estimated variance to define a distribution for each number of atlases. From these distribution estimates, we determined that the segmentation with 180 atlases outperformed (p<0.05, t-test) the segmentation with up to 160 atlases. We also determined that though the results did not show significant differences after 160 atlases (p<0.05; t-test), the variance results were still showing a trend downward and the mean Dice were trending upward (figure 4).

4.2. Voxel-Wise Variance Estimation with Joint Label Fusion

Since this is the first work considering voxel-wise segmentation, there is not a direct comparison available for the proposed method. In order to determine the optimal number of estimators, each of the ten targets was first considered individually. For each target and for each number of atlases up to 100, the number of voxels likely to change between iterations by parameterizing each voxel with a Gaussian distribution given the mean and variance estimates for its probability of being in the hippocampus. With that distribution, the cumulative density function of being below 0.5, the threshold for being called as hippocampus, was calculated, and then the likelihood of changing labels with a unique sample of atlases can be determined. Given these estimates for each number of atlases (Figure 5), we can see that with 100 atlases the percent of the voxels in the hippocampus that would change with a new random draw of labels is still decreasing, whereas the equivalent experiment using Monte Carlo estimation shows that there is no change with two atlas populations.

5. Discussion

The typical leave-one-out paradigm for determining the sufficient number of atlases in multi-atlas segmentation is flawed in that (1) it does not consider the samples in the evaluation set as individuals, (2) it does not model the covariance of the Monte Carlo segmentation process, and (3) it does not consider that two segmentations can have the same accuracy while having different boundaries of the image. For applications such as f-MRI correlations and DTI fiber tracking, the accuracy and consistency of the segmentation is vital to the success of the algorithms [82, 180]. Several techniques are available to estimate unknown distributions from a series of estimators. The most common technique is bootstrapping, where a sample of estimators are drawn from the pool of available estimators. These estimators are then fused together to form a consistent representation. Unfortunately, bootstrapping and related techniques are not applicable to label fusion techniques since many of the top-performing algorithms require an independence between the atlases. For instance, the joint label fusion (JLF) algorithm determines the covariance between the atlases. If atlases were duplicated in JLF, the algorithm would down-weight them proportionally by the number of duplicates of that atlas in the sample of atlases [84, 97, 121]. This would be the equivalent result to simply using the number of unique atlases in the atlas sample.

In this work, we have proposed a framework using Monte Carlo sampling to produce variance estimations more consistent with the true variance estimations. The proposed approach utilized pools of increasing numbers of atlases and the estimated percent similarity metric to fit a curve to the variance estimates and thus allowing us to estimate for variance values outside of the range of values which were empirically available in the data. These results were compared with a bootstrap approach and a pooled variance approach, both of which produce valid estimates of variance for the majority vote label fusion approach. The proposed approach produced

significantly more accurate results, and the proposed segmentation approach provides significantly different results from the standard Monte Carlo approach (Figure 4C). The Monte Carlo approach identified 136 as the optimal number of atlases for the segmentation, whereas the proposed segmentation determined that more than 150 atlases was needed for the segmentation approach. It is important to understand that reducing the variance is an important consideration, along with increasing the accuracy. In many cases, the cost of labeling more atlases may out-weigh the increases in accuracy and decreases in variance. Techniques like atlas selection may help account with these limitations, but such approaches are application dependent.

Finally, we extended the variance estimation framework to work on voxel-wise probabilistic segmentation results. In this work, we considered the differences in two segmentations from a random draw of atlases from an infinite population. This assesses if a segmentation has converged to a stable result on a voxel-wise basis instead of relying on summary statistics that may hide the underlying results.

Chapter VIII

Conclusions

1. Summary

When I began my dissertation, multi-atlas segmentation was a burgeoning field of research within the medical image processing community. Recently, several groups including the MASI lab at Vanderbilt had characterized approaches for accurate segmentation of various organs with multi-atlas segmentation. This dissertation has primarily focused on characterizing multi-atlas segmentation and expanding on previous theory to improve our understanding of the approaches available. Chapter II characterized an approach for segmentation when atlases did not have matching label sets associated with them. Chapter III described an algorithm for segmentation when the sequence of the target image does not match the sequence of the atlases. Chapters IV-VI present specialized segmentation approaches for particular regions of interest. Finally, Chapter VII describes an approach for evaluation of segmentation accuracy which is more appropriate than the standard and previously used approaches.

These approaches centered on improving our ability to characterize PD. The contributions focused on segmentation of structures of interest to studying the disease. In several chapters, my work focused directly on segmentation of regions of interest that are necessary for DBS surgery in PD. In other chapters, my work focused on theoretic contributions to multi-atlas segmentation. These contributions focused on improving our understanding of the number of atlases needed for a segmentation task.

2. Segmentation with Multiple Label Sets

We present a segmentation approach for multi-atlas segmentation allowing for incorporation of multiple label sets (Chapter II). This method expanded on the standard STAPLE framework and utilized a joint performance matrix characterization for all of the label sets simultaneously. This approach was then simplified to reduce the number of degrees of freedom, thus allowing for improved performance of the approach. These algorithms were first validated on a simulation against previous approaches, and the proposed algorithms outperformed the standard approaches. Then, the proposed algorithms were expanded to incorporate non-local correspondence and were applied to a real dataset on the human brain. The proposed approaches outperformed the previous approaches, in particular when there were few atlases of the "target" protocol available.

3. Segmentation with Multiple Imaging Sequences

We present a segmentation approach optimized to decrease variance between imaging sequences (Chapter III). We identified a bias present when different imaging sequences were used between the target and atlas images. We proposed synthesizing atlases that more similarly match the target scan. This approach incorporates atlases with available biological parameters maps, namely T1- and T2- relaxation and a PD map). We tested this approach on a population of target images with varying sequences and we showed that our proposed algorithm significantly decreased the bias between the sequences. We further compared our ability to differentiate healthy and autistic subjects from the ABIDE study using segmentation results from the proposed and standard segmentation approaches. We showed that by using our synthetic approach we gained some power in differentiating the different populations.

4. Segmentation of Specialized Anatomy

We present three segmentation approaches for different anatomic structures. First (Chapter IV), we present an approach for automated segmentation of the hippocampus and amygdala. We use nearly 200 atlases and a "reduced field of view" segmentation to efficiently register these scans to the target image and automatically segment the target image. Our proposed approach outperformed other segmentation approaches and using the full atlas set outperformed using just 30 atlases. Second (Chapter V), we present an approach for segmentation of the sub-cortical grey matter. In this work, we utilized specialized imaging sequences to improve the segmentation accuracy. By selectively incorporating T1-weighted MPRAGE scans and F-GATIR scans we were able to improve the accuracy of the segmentation approach. Third (Chapter VI), we present a segmentation approach for segmentation of the cerebellum. The cerebellum is an interesting structure to study with imaging because older and diseased subjects show a higher degree of anatomic differentiability. We propose a segmentation approach, non-local SIMPLE, which is a non-local segmentation approach where patches are treated as functional units in an atlas selection framework instead of the standard atlas based approach. This approach outperformed all other approaches on a heavily diseased population, but on a healthier population more standard approaches performed comparably or better than non-local SIMPLE.

5. Estimation of Variance in Multi-Atlas Segmentation

In order to properly determine the number of atlases needed for a segmentation, an estimate of the variance of a given subject's results is necessary. Furthermore, a segmentation should seek to produce results with as low of variation as possible with two unique populations of atlases. Current approaches for estimation of the number of atlases for segmentation and the accuracy of segmentation approaches do not consider this difference. We propose an algorithm for proper

calculation of variance when it is impossible to achieve truly distinct populations of atlases. We use this approach to first properly estimate the performance of a segmentation result using Dice Similarity Coefficient. Second, we use this approach to calculate the voxel-wise variance and determine the number of atlases needed minimize the number of voxels changing between two atlas populations.

6. Summary of Contributions

The final contributions of this dissertation are summarized as follows

- We developed several segmentation techniques which advance our understanding of statistical fusion in several contexts. In particular this includes the case where we have multiple labeling protocols which have similar or complimentary protocols and we wish to fuse them together. This also includes the case where we treat a patch derived from an atlas as the statistical unit in segmentation as opposed to the atlas itself.
- We proposed a segmentation approach which minimizes the variability between T1-weighted imaging sequences. This addresses the issue where there is a significant bias in segmentation results as a function of the imaging sequence instead of any true variance in the data
- We developed a segmentation algorithm for the temporal lobe, particularly the amygdala and hippocampus. This approach is both efficient and accurate for segmentation of two of the most studied structures in the human brain.
- We developed a segmentation algorithm for accurate segmentation of the human subcortex. This approach utilized specialized imaging sequences, targeted at these structures

- We developed a segmentation approach for segmentation of the cerebellum when the anatomy presents in highly variable patterns.
- We developed an algorithm for assessing the importance of different imaging modalities in multi-atlas segmentation. Much of imaging research is becoming multi-modal as structures and metrics of interest and it is necessary to have techniques to understand the importance and necessity of them.

7. Impact on PD

This dissertation focused on PD, a debilitating disease that effects patients and their families. The advancements in this work provide new in avenues for differentiating PD from other diseases, tracking the progression of PD, and understanding the mechanisms underlying the disease. The various segmentation approaches and theoretic advancements can also be translated to other conditions of interest, and thus may provide inference in other circumstances. My opportunity to work on PD has been a pleasure. Being able to contribute to a variety of research projects and work with so many passionate people has made my time working on PD a pleasure.

References

1. Shulman, J.M., De Jager, P.L., Feany, M.B.: Parkinson's disease: genetics and pathogenesis. Annual Review of Pathology: Mechanisms of Disease 6, 193-222 (2011)

2. Jankovic, J.: Parkinson’s disease: clinical features and diagnosis. Journal of Neurology, Neurosurgery & Psychiatry 79, 368-376 (2008)

3. Caballol, N., Martí, M.J., Tolosa, E.: Cognitive dysfunction and dementia in Parkinson disease. Movement Disorders 22, S358-S366 (2007)

4. Parker, K.L., Lamichhane, D., Caetano, M.S., Narayanan, N.S.: Executive dysfunction in Parkinson’s disease and timing deficits. Front Integr Neurosci 7, 75.10 (2013)

5. Vos, T., Barber, R.M., Bell, B., Bertozzi-Villa, A., Biryukov, S., Bolliger, I., Charlson, F., Davis, A., Degenhardt, L., Dicker, D.: Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. The Lancet 386, 743-800 (2015)

6. Braak, H., Del Tredici, K., Rüb, U., de Vos, R.A., Steur, E.N.J., Braak, E.: Staging of brain pathology related to sporadic Parkinson’s disease. Neurobiology of aging 24, 197-211 (2003)

7. Calne, D., Snow, B., Lee, C.: Criteria for diagnosing Parkinson's disease. Annals of neurology 32, S125-S127 (1992)

8. Conditions, N.C.C.f.C.: Diagnosing Parkinson’s disease. (2006)

9. Dickson, D.W., Feany, M., Yen, S.-H., Mattiace, L., Davies, P.: Cytoskeletal pathology in non-Alzheimer degenerative dementia: new lesions in diffuse Lewy body disease, Pick’s disease, and corticobasal degeneration. Springer (1996)

10. Worth, P.F.: How to treat Parkinson's disease in 2013. Clinical Medicine 13, 93-96 (2013)

11. Jenner, P., Dexter, D., Sian, J., Schapira, A., Marsden, C.: Oxidative stress as a cause of nigral cell death in Parkinson's disease and incidental Lewy body disease. Annals of Neurology 32, S82-S87 (1992)

12. Aggleton, E.J., Everitt, B.J., Cardinal, R.N., Hall, J.: The amygdala: a functional analysis. (2000)

13. Remy, P., Doder, M., Lees, A., Turjanski, N., Brooks, D.: Depression in Parkinson's disease: loss of dopamine and noradrenaline innervation in the limbic system. Brain 128, 1314-1322 (2005)

14. Pascual-Leone, A., Grafman, J., Clark, K., Stewart, M., Massaquoi, S., Lou, J.S., Hallett, M.: Procedural learning in Parkinson's disease and cerebellar degeneration. Annals of neurology 34, 594-602 (1993)

15. Alexander, G.E., Crutcher, M.D., DeLong, M.R.: Basal ganglia-thalamocortical circuits: parallel substrates for motor, oculomotor," prefrontal" and" limbic" functions. Progress in brain research 85, 119-146 (1989)

16. Alexander, G.E., Crutcher, M.D.: Functional architecture of basal ganglia circuits: neural substrates of parallel processing. Trends in neurosciences 13, 266-271 (1990)

17. Obeso, J.A., Rodríguez-Oroz, M.C., Benitez-Temino, B., Blesa, F.J., Guridi, J., Marin, C., Rodriguez, M.: Functional organization of the basal ganglia: therapeutic implications for Parkinson's disease. Movement Disorders 23, S548-S559 (2008)

18. Yager, L., Garcia, A., Wunsch, A., Ferguson, S.: The ins and outs of the striatum: Role in drug addiction. Neuroscience 301, 529-541 (2015)

19. Fox, C., Andrade, A., Lu, Q.I., Rafols, J.: The primate globus pallidus: a Golgi and electron microscopic study. Journal fur Hirnforschung 15, 75-93 (1973)

20. Surmeier, D.J., Mercer, J.N., Chan, C.S.: Autonomous pacemakers in the basal ganglia: who needs excitatory synapses anyway? Current opinion in neurobiology 15, 312-318 (2005)

21. Frank, M.J., Samanta, J., Moustafa, A.A., Sherman, S.J.: Hold your horses: impulsivity, deep brain stimulation, and medication in parkinsonism. Science 318, 1309-1312 (2007)

22. Deniau, J., Kitai, S., Donoghue, J., Grofova, I.: Neuronal interactions in the substantia nigra pars reticulata through axon collaterals of the projection neurons. Experimental brain research 47, 105-113 (1982)

23. Kish, S.J., Shannak, K., Hornykiewicz, O.: Uneven pattern of dopamine loss in the striatum of patients with idiopathic Parkinson's disease. New England Journal of Medicine 318, 876-880 (1988)

24. Chan, C.S., Glajch, K.E., Gertler, T.S., Guzman, J.N., Mercer, J.N., Lewis, A.S., Goldberg, A.B., Tkatch, T., Shigemoto, R., Fleming, S.M.: HCN channelopathy in external globus pallidus neurons in models of Parkinson's disease. Nature neuroscience 14, 85-92 (2011)

25. Stefani, A., Stanzione, P., Bassi, A., Mazzone, P., Vangelista, T., Bernardi, G.: Effects of increasing doses of apomorphine during stereotaxic neurosurgery in Parkinson's disease: clinical score and internal globus pallidus activity. Journal of neural transmission 104, 895-904 (1997)

26. Rodriguez-Oroz, M.C., Rodriguez, M., Guridi, J., Mewes, K., Chockkman, V., Vitek, J., DeLong, M.R., Obeso, J.A.: The subthalamic nucleus in Parkinson's disease: somatotopic organization and physiological characteristics. Brain 124, 1777-1790 (2001)

27. Fearnley, J.M., Lees, A.J.: Ageing and Parkinson's disease: substantia nigra regional selectivity. Brain 114, 2283-2301 (1991)

28. Behrens, T., Johansen-Berg, H., Woolrich, M., Smith, S., Wheeler-Kingshott, C., Boulby, P., Barker, G., Sillery, E., Sheehan, K., Ciccarelli, O.: Non-invasive mapping of connections

between human thalamus and cortex using diffusion imaging. Nature neuroscience 6, 750-757 (2003)

29. Steriade, M., Llinás, R.R.: The functional states of the thalamus and the associated neuronal interplay. Physiological reviews 68, 649-742 (1988)

30. Herrero, M.-T., Barcia, C., Navarro, J.: Functional anatomy of thalamus and basal ganglia. Child's Nervous System 18, 386-404 (2002)

31. Li, X.B., Inoue, T., Nakagawa, S., Koyama, T.: Effect of mediodorsal thalamic nucleus lesion on contextual fear conditioning in rats. Brain research 1008, 261-272 (2004)

32. Abitz, M., Nielsen, R.D., Jones, E.G., Laursen, H., Graem, N., Pakkenberg, B.: Excess of neurons in the human newborn mediodorsal thalamus compared with that of the adult. Cerebral Cortex 17, 2573-2578 (2007)

33. Blomqvist, A., Zhang, E.-T., Craig, A.: Cytoarchitectonic and immunohistochemical characterization of a specific pain and temperature relay, the posterior portion of the ventral medial nucleus, in the human thalamus. Brain 123, 601-619 (2000)

34. Rose, J.E., Woolsey, C.N.: Structure and relations of limbic cortex and anterior thalamic nuclei in rabbit and cat. Journal of Comparative Neurology 89, 279-347 (1948)

35. van Groen, T., Kadish, I., Wyss, J.M.: Role of the anterodorsal and anteroventral nuclei of the thalamus in spatial memory in the rat. Behavioural brain research 132, 19-28 (2002)

36. Jones, E., Powell, T.: An analysis of the posterior group of thalamic nuclei on the basis of its afferent connections. Journal of Comparative Neurology 143, 185-215 (1971)

37. O'Connor, D.H., Fukui, M.M., Pinsk, M.A., Kastner, S.: Attention modulates responses in the human lateral geniculate nucleus. Nature neuroscience 5, 1203-1209 (2002)

38. Heath, C., Jones, E.: An experimental study of ascending connections from the posterior group of thalamic nuclei in the cat. Journal of Comparative Neurology 141, 397-425 (1971)

39. Deiber, M.-P., Pollak, P., Passingham, R., Landais, P., Gervason, C., Cinotti, L., Friston, K., Frackowiak, R., Mauguière, F., Benabid, A.L.: Thalamic stimulation and suppression of parkinsonian tremor. Brain 116, 267-279 (1993)

40. Bacci, J.-J., Kachidian, P., Kerkerian-Le Goff, L., Salin, P.: Intralaminar thalamic nuclei lesions: widespread impact on dopamine denervation-mediated cellular defects in the rat basal ganglia. Journal of Neuropathology & Experimental Neurology 63, 20-31 (2004)

41. Ferraye, M., Debû, B., Fraix, V., Goetz, L., Ardouin, C., Yelnik, J., Henry-Lagrange, C., Seigneuret, E., Piallat, B., Krack, P.: Effects of pedunculopontine nucleus area stimulation on gait disorders in Parkinson's disease. Brain 133, 205-214 (2010)

42. Eccles, J.C.: The cerebellum as a neuronal machine. Springer Science & Business Media (2013)

43. Holmes, G.: The cerebellum of man. Brain 62, 1-30 (1939)

44. Schmahmann, J.D., Caplan, D.: Cognition, emotion and the cerebellum. Brain 129, 290-292 (2006)

45. Asanuma, C., Thach, W., Jones, E.: Distribution of cerebellar terminations and their relation to other afferent terminations in the ventral lateral thalamic region of the monkey. Brain Research Reviews 5, 237-265 (1983)

46. Herrup and, K., Kuemerle, B.: The compartmentalization of the cerebellum. Annual review of neuroscience 20, 61-90 (1997)

47. Wu, T., Hallett, M.: The cerebellum in Parkinson’s disease. Brain 136, 696-709 (2013)

48. Gilman, S., Koeppe, R., Nan, B., Wang, C.-N., Wang, X., Junck, L., Chervin, R., Consens, F., Bhaumik, A.: Cerebral cortical and subcortical cholinergic deficits in parkinsonian syndromes. Neurology 74, 1416-1423 (2010)

49. Nishio, Y., Hirayama, K., Takeda, A., Hosokai, Y., Ishioka, T., Suzuki, K., Itoyama, Y., Takahashi, S., Mori, E.: Corticolimbic gray matter loss in Parkinson's disease without dementia. European journal of neurology 17, 1090-1097 (2010)

50. Mogenson, G.J., Jones, D.L., Yim, C.Y.: From motivation to action: functional interface between the limbic system and the motor system. Progress in neurobiology 14, 69-97 (1980)

51. Isaacson, R.L.: The structure of the limbic system. The Limbic System, pp. 1-60. Springer (1982)

52. Gellhorn, E., Loofbourrow, G.N.: Emotions and emotional disorders: A neurophysiological study. (1963)

53. Cassone, V.M., Speh, J.C., Card, J.P., Moore, R.Y.: Comparative anatomy of the mammalian hypothalamic suprachiasmatic nucleus. Journal of biological rhythms 3, 71-91 (1988)

54. Tulving, E., Markowitsch, H.J.: Episodic and declarative memory: role of the hippocampus. Hippocampus 8, 198-204 (1998)

55. Freund, T.F., Buzsáki, G.: Interneurons of the hippocampus. Hippocampus 6, 347-470 (1996)

56. Amaral, D.G., Capitanio, J.P., Jourdain, M., Mason, W.A., Mendoza, S.P., Prather, M.: The amygdala. Neuropsychologia 41, 235-240 (2003)

57. Saddoris, M.P., Sugam, J.A., Stuber, G.D., Witten, I.B., Deisseroth, K., Carelli, R.M.: Mesolimbic dopamine dynamically tracks, and is causally linked to, discrete aspects of value-based decision making. Biological psychiatry 77, 903-911 (2015)

58. Calipari, E.S., Bagot, R.C., Purushothaman, I., Davidson, T.J., Yorgason, J.T., Peña, C.J., Walker, D.M., Pirpinias, S.T., Guise, K.G., Ramakrishnan, C.: In vivo imaging identifies temporal signature of D1 and D2 medium spiny neurons in cocaine reward. Proceedings of the National Academy of Sciences 113, 2726-2731 (2016)

59. Langston, J.W., Forno, L.S.: The hypothalamus in Parkinson disease. Annals of neurology 3, 129-133 (1978)

60. Palacios, N., Gao, X., McCullough, M.L., Jacobs, E.J., Patel, A.V., Mayo, T., Schwarzschild, M.A., Ascherio, A.: Obesity, diabetes, and risk of Parkinson's disease. Movement Disorders 26, 2253-2259 (2011)

61. Scatton, B., Javoy-Agid, F., Rouquier, L., Dubois, B., Agid, Y.: Reduction of cortical dopamine, noradrenaline, serotonin and their metabolites in Parkinson's disease. Brain research 275, 321-328 (1983)

62. Otmakhova, N., Duzel, E., Deutch, A.Y., Lisman, J.: The hippocampal-VTA loop: the role of novelty and motivation in controlling the entry of information into long-term memory. Intrinsically motivated learning in natural and artificial systems, pp. 235-254. Springer (2013)

63. Braak, H., Braak, E., Yilmazer, D., de Vos, R.A., Jansen, E.N., Bohl, J., Jellinger, K.: Amygdala pathology in Parkinson's disease. Acta neuropathologica 88, 493-500 (1994)

64. Harding, A., Broe, G., Halliday, G.: Visual hallucinations in Lewy body disease relate to Lewy bodies in the temporal lobe. Brain 125, 391-403 (2002)

65. Cools, R., Lewis, S.J., Clark, L., Barker, R.A., Robbins, T.W.: L-DOPA disrupts activity in the nucleus accumbens during reversal learning in Parkinson's disease. Neuropsychopharmacology 32, 180-189 (2007)

66. Münchau, A., Bhatia, K.: Pharmacological treatment of Parkinson's disease. Postgraduate medical journal 76, 602-610 (2000)

67. Kahn, E., D'Haese, P.-F., Dawant, B., Allen, L., Kao, C., Charles, P.D., Konrad, P.: Deep brain stimulation in early stage Parkinson's disease: operative experience from a prospective randomised clinical trial. Journal of Neurology, Neurosurgery & Psychiatry jnnp-2011-300008 (2011)

68. Dunne, K., Sullivan, K., Kernohan, G.: Palliative care for patients with cancer: district nurses' experiences. Journal of advanced nursing 50, 372-380 (2005)

69. Pham, D.L., Xu, C., Prince, J.L.: Current methods in medical image segmentation 1. Annual review of biomedical engineering 2, 315-337 (2000)

70. Iglesias, J.E., Sabuncu, M.R.: Multi-atlas segmentation of biomedical images: a survey. Medical image analysis 24, 205-219 (2015)

71. Gee, J.C., Reivich, M., Bajcsy, R.: Elastically deforming 3D atlas to match anatomical brain images. Journal of computer assisted tomography 17, 225-236 (1993)

72. Fischl, B.: FreeSurfer. Neuroimage 62, 774-781 (2012)

73. Deichmann, R., Good, C., Josephs, O., Ashburner, J., Turner, R.: Optimization of 3-D MP-RAGE sequences for structural brain imaging. Neuroimage 12, 112-127 (2000)

74. Coleman, G.B., Andrews, H.C.: Image segmentation by clustering. Proceedings of the IEEE 67, 773-785 (1979)

75. Liang, Z., MacFall, J.R., Harrington, D.P.: Parameter estimation and tissue segmentation from multispectral MR images. IEEE transactions on medical imaging 13, 441-449 (1994)

76. Li, S.Z.: Markov random field modeling in computer vision. Springer Science & Business Media (2012)

77. Rajapakse, J.C., Giedd, J.N., Rapoport, J.L.: Statistical approach to segmentation of single-channel cerebral MR images. IEEE transactions on medical imaging 16, 176-186 (1997)

78. Held, K., Kops, E.R., Krause, B.J., Wells, W.M., Kikinis, R., Muller-Gartner, H.-W.: Markov random field segmentation of brain MR images. IEEE transactions on medical imaging 16, 878-886 (1997)

79. McInerney, T., Terzopoulos, D.: Deformable models in medical image analysis: a survey. Medical image analysis 1, 91-108 (1996)

80. Xu, C., Prince, J.L.: Snakes, shapes, and gradient vector flow. IEEE Transactions on image processing 7, 359-369 (1998)

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

82. Mori, S., van Zijl, P.: Fiber tracking: principles and strategies–a technical review. NMR in Biomedicine 15, 468-480 (2002)

83. Yendiki, A., Panneck, P., Srinivasan, P., Stevens, A., Zöllei, L., Augustinack, J., Wang, R., Salat, D., Ehrlich, S., Behrens, T.: Automated probabilistic reconstruction of white-matter pathways in health and disease using an atlas of the underlying anatomy. Front. Neuroinf. 5, 12-23 (2011)

84. Sabuncu, M.R., Yeo, B.T., Van Leemput, K., Fischl, B., Golland, P.: A generative model for image segmentation based on label fusion. IEEE transactions on medical imaging 29, 1714-1729 (2010)

85. Menke, R.A., Scholz, J., Miller, K.L., Deoni, S., Jbabdi, S., Matthews, P.M., Zarei, M.: MRI characteristics of the substantia nigra in Parkinson's disease: a combined quantitative T1 and DTI study. Neuroimage 47, 435-441 (2009)

86. Ourselin, S., Roche, A., Subsol, G., Pennec, X., Ayache, N.: Reconstructing a 3D structure from serial histological sections. Image Vision Comput 19, 25-31 (2001)

87. Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C.: A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 54, 2033-2044 (2011)

88. Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med Image Anal 12, 26-41 (2008)

89. Iglesias, J.E., Sabuncu, M.R.: Multi-atlas segmentation of biomedical images: A survey. Med Image Anal 24, 205-219 (2015)

90. Artaechevarria, X., Munoz-Barrutia, A., Ortiz-de-Solórzano, C.: Combination strategies in multi-atlas image segmentation: Application to brain MR data. IEEE transactions on medical imaging 28, 1266-1277 (2009)

91. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE transactions on medical imaging 23, 903-921 (2004)

92. Asman, A.J., Landman, B.A.: Formulating spatially varying performance in the statistical fusion framework. IEEE transactions on medical imaging 31, 1326-1336 (2012)

93. Asman, A.J., Landman, B.A.: Hierarchical performance estimation in the statistical label fusion framework. Med Image Anal 18, 1070-1081 (2014)

94. Manjón, J.V., Carbonell-Caballero, J., Lull, J.J., García-Martí, G., Martí-Bonmatí, L., Robles, M.: MRI denoising using non-local means. Medical image analysis 12, 514-523 (2008)

95. Buades, A., Coll, B., Morel, J.-M.: A non-local algorithm for image denoising. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, pp. 60-65. IEEE, (Year)

96. Coupe, P., Manjon, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L.: Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. Neuroimage 54, 940-954 (2011)

97. Asman, A.J., Landman, B.A.: Non-local statistical label fusion for multi-atlas segmentation. Med Image Anal 17, 194-208 (2013)

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

99. Wang, H., Das, S.R., Suh, J.W., Altinay, M., Pluta, J., Craige, C., Avants, B., Yushkevich, P.A., Initiative, A.s.D.N.: A learning-based wrapper method to correct systematic errors in automatic image segmentation: consistently improved performance in hippocampus, cortex and brain segmentation. NeuroImage 55, 968-985 (2011)

100. Xu, Z., Conrad, B.N., Baucom, R.B., Smith, S.A., Poulose, B.K., Landman, B.A.: Abdomen and spinal cord segmentation with augmented active shape models. Journal of Medical Imaging 3, 036002-036002 (2016)

101. Yang, Z., Ye, C., Bogovic, J.A., Carass, A., Jedynak, B.M., Ying, S.H., Prince, J.L.: Automated cerebellar lobule segmentation with application to cerebellar structural analysis in cerebellar disease. NeuroImage 127, 435-444 (2016)

102. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology 26, 297-302 (1945)

103. Hausdorff, F.: Mengenlehre. Walter de Gruyter Berlin (1927)

104. Wolz, R., Chu, C., Misawa, K., Fujiwara, M., Mori, K., Rueckert, D.: Automated abdominal multi-organ segmentation with subject-specific atlas generation. IEEE transactions on medical imaging 32, 1723-1730 (2013)

105. Langerak, T.R., van der Heide, U.A., Kotte, A.N., Viergever, M.A., van Vulpen, M., Pluim, J.P.: Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). IEEE transactions on medical imaging 29, 2000-2008 (2010)

106. Konrad, C., Ukas, T., Nebel, C., Arolt, V., Toga, A.W., Narr, K.L.: Defining the human hippocampus in cerebral magnetic resonance images--an overview of current segmentation protocols. Neuroimage 47, 1185-1195 (2009)

107. Iglesias, J.E., Sabuncu, M.R., Aganj, I., Bhatt, P., Casillas, C., Salat, D., Boxer, A., Fischl, B., Van Leemput, K.: An algorithm for optimal fusion of atlases with different labeling protocols. Neuroimage (2014)

108. Oishi, K., Faria, A., Jiang, H., Li, X., Akhter, K., Zhang, J., Hsu, J.T., Miller, M.I., van Zijl, P.C., Albert, M., Lyketsos, C.G., Woods, R., Toga, A.W., Pike, G.B., Rosa-Neto, P., Evans, A., Mazziotta, J., Mori, S.: Atlas-based whole brain white matter analysis using large deformation diffeomorphic metric mapping: application to normal elderly and Alzheimer's disease participants. Neuroimage 46, 486-499 (2009)

109. Iglesias, J.E., Sabuncu, M.R.: Multi-Atlas Segmentation of Biomedical Images: A Survey. arXiv preprint arXiv:1412.3421 (2014)

110. Fox, N.C., Freeborough, P.A.: Brain atrophy progression measured from registered serial MRI: validation and application to Alzheimer's disease. Journal of Magnetic Resonance Imaging 7, 1069-1075 (1997)

111. Greicius, M.D., Supekar, K., Menon, V., Dougherty, R.F.: Resting-state functional connectivity reflects structural connectivity in the default mode network. Cerebral cortex 19, 72-78 (2009)

112. Behrens, T., Berg, H.J., Jbabdi, S., Rushworth, M., Woolrich, M.: Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? Neuroimage 34, 144-155 (2007)

113. Plassard, A.J., Harrigan, R.L., Newton, A.T., Rane, S., Pallavaram, S., D'Haese, P.F., Dawant, B.M., Claassen, D.O., Landman, B.A.: On the fallacy of quantitative segmentation for T1-weighted MRI. In: SPIE Medical Imaging, pp. 978416-978416-978417. International Society for Optics and Photonics, (Year)

114. Jovicich, J., Czanner, S., Han, X., Salat, D., van der Kouwe, A., Quinn, B., Pacheco, J., Albert, M., Killiany, R., Blacker, D.: MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. Neuroimage 46, 177-192 (2009)

115. Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L Whitwell, J., Ward, C.: The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. Journal of Magnetic Resonance Imaging 27, 685-691 (2008)

116. Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M.: The autism brain imaging data exchange: towards

a large-scale evaluation of the intrinsic brain architecture in autism. Molecular psychiatry 19, 659-667 (2014)

117. Tustison, N.J., Cook, P.A., Klein, A., Song, G., Das, S.R., Duda, J.T., Kandel, B.M., van Strien, N., Stone, J.R., Gee, J.C.: Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. Neuroimage 99, 166-179 (2014)

118. Nielsen, J.A., Zielinski, B.A., Fletcher, P.T., Alexander, A.L., Lange, N., Bigler, E.D., Lainhart, J.E., Anderson, J.S.: Multisite functional connectivity MRI classification of autism: ABIDE results. (2013)

119. Jog, A., Carass, A., Roy, S., Pham, D.L., Prince, J.L.: MR image synthesis by contrast learning on neighborhood ensembles. Medical image analysis 24, 63-76 (2015)

120. Lorenzi, M., Ziegler, G., Alexander, D.C., Ourselin, S.: Efficient Gaussian process-based modelling and prediction of image time series. In: Information Processing in Medical Imaging, pp. 626-637. Springer, (Year)

121. Wang, H., Yushkevich, P.A.: Groupwise segmentation with multi-atlas joint label fusion. Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention 16, 711-718 (2013)

122. Reich, D.S., Smith, S.A., Zackowski, K.M., Gordon-Lipkin, E.M., Jones, C.K., Farrell, J.A., Mori, S., van Zijl, P.C., Calabresi, P.A.: Multiparametric magnetic resonance imaging analysis of the corticospinal tract in multiple sclerosis. Neuroimage 38, 271-279 (2007)

123. Bernstein, M.A., King, K.F., Zhou, X.J.: Handbook of MRI pulse sequences. Elsevier (2004)

124. Wansapura, J.P., Holland, S.K., Dunn, R.S., Ball, W.S.: NMR relaxation times in the human brain at 3.0 tesla. J Magn Reson Imaging 9, 531-538 (1999)

125. Landman, B.A., Huang, A.J., Gifford, A., Vikram, D.S., Lim, I.A.L., Farrell, J.A., Bogovic, J.A., Hua, J., Chen, M., Jarso, S.: Multi-parametric neuroimaging reproducibility: a 3-T resource study. Neuroimage 54, 2854-2866 (2011)

126. Krishnapuram, B., Carin, L., Figueiredo, M.A., Hartemink, A.J.: Sparse multinomial logistic regression: Fast algorithms and generalization bounds. Pattern Analysis and Machine Intelligence, IEEE Transactions on 27, 957-968 (2005)

127. Eichenbaum, H.: Hippocampus: cognitive processes and neural representations that underlie declarative memory. Neuron 44, 109-120 (2004)

128. Janak, P.H., Tye, K.M.: From circuits to behaviour in the amygdala. Nature 517, 284-292 (2015)

129. Small, S.A., Schobel, S.A., Buxton, R.B., Witter, M.P., Barnes, C.A.: A pathophysiological framework of hippocampal dysfunction in ageing and disease. Nature Reviews Neuroscience 12, 585-601 (2011)

130. Andersen, P.: The hippocampus book. Oxford university press (2007)

131. Fanselow, M.S., Dong, H.-W.: Are the dorsal and ventral hippocampus functionally distinct structures? Neuron 65, 7-19 (2010)

132. Yassa, M.A., Stark, C.E.: Pattern separation in the hippocampus. Trends in neurosciences 34, 515-525 (2011)

133. Aggleton, J.P.: Multiple anatomical systems embedded within the primate medial temporal lobe: implications for hippocampal function. Neuroscience & Biobehavioral Reviews 36, 1579-1596 (2012)

134. Lisman, J.E., Grace, A.A.: The hippocampal-VTA loop: controlling the entry of information into long-term memort. Neuron 46, 703-713 (2005)

135. Poppenk, J., Evensmoen, H.R., Moscovitch, M., Nadel, L.: Long-axis specialization of the human hippocampus. Trends in cognitive sciences 17, 230-240 (2013)

136. Strange, B.A., Witter, M.P., Lein, E.S., Moser, E.I.: Functional organization of the hippocampal longitudinal axis. Nature Reviews Neuroscience 15, 655-669 (2014)

137. Zeidman, P., Maguire, E.A.: Anterior hippocampus: the anatomy of perception, imagination and episodic memory. Nature Reviews Neuroscience 17, 173-182 (2016)

138. Amunts, K., Kedo, O., Kindler, M., Pieperhoff, P., Mohlberg, H., Shah, N., Habel, U., Schneider, F., Zilles, K.: Cytoarchitectonic mapping of the human amygdala, hippocampal region and entorhinal cortex: intersubject variability and probability maps. Anatomy and embryology 210, 343-352 (2005)

139. Adolphs, R., Tranel, D., Damasio, H., Damasio, A.: Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. Nature 372, 669-672 (1994)

140. Gee, D.G., Humphreys, K.L., Flannery, J., Goff, B., Telzer, E.H., Shapiro, M., Hare, T.A., Bookheimer, S.Y., Tottenham, N.: A developmental shift from positive to negative connectivity in human amygdala–prefrontal circuitry. The Journal of Neuroscience 33, 4584-4593 (2013)

141. Rutishauser, U., Mamelak, A.N., Adolphs, R.: The primate amygdala in social perception–insights from electrophysiological recordings and stimulation. Trends in neurosciences 38, 295-306 (2015)

142. Goodkind, M., Eickhoff, S.B., Oathes, D.J., Jiang, Y., Chang, A., Jones-Hagata, L.B., Ortega, B.N., Zaiko, Y.V., Roach, E.L., Korgaonkar, M.S.: Identification of a common neurobiological substrate for mental illness. JAMA psychiatry 72, 305-315 (2015)

143. Okada, N., Fukunaga, M., Yamashita, F., Koshiyama, D., Yamamori, H., Ohi, K., Yasuda, Y., Fujimoto, M., Watanabe, Y., Yahata, N.: Abnormal asymmetries in subcortical brain volume in schizophrenia. Molecular psychiatry (2016)

144. Schneider, F., Weiss, U., Kessler, C., Salloum, J., Posse, S., Grodd, W., Müller-Gärtner, H.: Differential amygdala activation in schizophrenia during sadness. Schizophrenia research 34, 133-142 (1998)

145. Giedd, J.N., Vaituzis, A.C., Hamburger, S.D., Lange, N., Rajapakse, J.C., Kaysen, D., Vauss, Y.C., Rapoport, J.L.: Quantitative MRI of the temporal lobe, amygdala, and hippocampus in normal human development: ages 4–18 years. Journal of Comparative Neurology 366, 223-230 (1996)

146. Patenaude, B., Smith, S.M., Kennedy, D.N., Jenkinson, M.: A Bayesian model of shape and appearance for subcortical brain segmentation. Neuroimage 56, 907-922 (2011)

147. Yushkevich, P.A., Wang, H., Pluta, J., Das, S.R., Craige, C., Avants, B.B., Weiner, M.W., Mueller, S.: Nearly automatic segmentation of hippocampal subfields in in vivo focal T2-weighted MRI. Neuroimage 53, 1208-1224 (2010)

148. Harrigan, R.L., Panda, S., Asman, A.J., Nelson, K.M., Chaganti, S., DeLisi, M.P., Yvernault, B.C., Smith, S.A., Galloway, R.L., Mawn, L.A.: Robust optic nerve segmentation on clinically acquired computed tomography. Journal of Medical Imaging 1, 034006-034006 (2014)

149. Panda, S., Asman, A.J., Khare, S.P., Thompson, L., Mawn, L.A., Smith, S.A., Landman, B.A.: Evaluation of multiatlas label fusion for in vivo magnetic resonance imaging orbital segmentation. Journal of Medical Imaging 1, 024002-024002 (2014)

150. Xu, Z., Burke, R.P., Lee, C.P., Baucom, R.B., Poulose, B.K., Abramson, R.G., Landman, B.A.: Efficient multi-atlas abdominal segmentation on clinically acquired CT with SIMPLE context learning. Medical image analysis 24, 18-27 (2015)

151. First, M.B., Spitzer, R.L., Gibbon, M., Williams, J.B.: Structured clinical interview for DSM-IV-TR axis I disorders, research version, patient edition. SCID-I/P (2002)

152. Woolard, A.A., Heckers, S.: Anatomical and functional correlates of human hippocampal volume asymmetry. Psychiatry Research: Neuroimaging 201, 48-53 (2012)

153. Pruessner, J., Li, L., Serles, W., Pruessner, M., Collins, D., Kabani, N., Lupien, S., Evans, A.: Volumetry of hippocampus and amygdala with high-resolution MRI and three-dimensional analysis software: minimizing the discrepancies between laboratories. Cerebral cortex 10, 433-442 (2000)

154. Amaral, D.G., Witter, M.: The three-dimensional organization of the hippocampal formation: a review of anatomical data. Neuroscience 31, 571-591 (1989)

155. Boccardi, M., Ganzola, R., Bocchetta, M., Pievani, M., Redolfi, A., Bartzokis, G., Camicioli, R., Csernansky, J.G., De Leon, M.J., Detoledo-Morrell, L.: Survey of protocols for the manual segmentation of the hippocampus: preparatory steps towards a joint EADC-ADNI harmonized protocol. Journal of Alzheimer's Disease 26, 61-75 (2011)

156. Fine, E.J., Ionita, C.C., Lohr, L.: The history of the development of the cerebellar examination. Semin Neurol 22, 375-384 (2002)

157. Timmann, D., Daum, I.: Cerebellar contributions to cognitive functions: a progress report after two decades of research. Cerebellum 6, 159-162 (2007)

158. Strick, P.L., Dum, R.P., Fiez, J.A.: Cerebellum and nonmotor function. Annu Rev Neurosci 32, 413-434 (2009)

159. van der Lijn, F., De Bruijne, M., Hoogendam, Y.Y., Klein, S., Hameeteman, R., Breteler, M., Niessen, W.J.: Cerebellum segmentation in MRI using atlas registration and local multi-scale image descriptors. In: Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on, pp. 221-224. IEEE, (Year)

160. Saeed, N., Puri, B.: Cerebellum segmentation employing texture properties and knowledge based image processing: applied to normal adult controls and patients. Magnetic resonance imaging 20, 425-429 (2002)

161. Powell, S., Magnotta, V.A., Johnson, H., Jammalamadaka, V.K., Pierson, R., Andreasen, N.C.: Registration and machine learning-based automated segmentation of subcortical and cerebellar brain structures. Neuroimage 39, 238-247 (2008)

162. Diedrichsen, J., Balsters, J.H., Flavell, J., Cussans, E., Ramnani, N.: A probabilistic MR atlas of the human cerebellum. Neuroimage 46, 39-46 (2009)

163. Xu, Z., Asman, A.J., Shanahan, P.L., Abramson, R.G., Landman, B.A.: SIMPLE is a good idea (and better with context learning). Med Image Comput Comput Assist Interv 17, 364-371 (2014)

164. Agarwal, M., Hendriks, E., Stoel, B., Bakker, M., Reiber, J., Staring, M.: Local SIMPLE multi atlas-based segmentation applied to lung lobe detection on chest CT. In: SPIE Medical Imaging, pp. 831410-831410-831417. International Society for Optics and Photonics, (Year)

165. Yang, Z., Bogovic, J.A., Ye, C., Ying, S.H., Prince, J.L.: Automated Cerebellar Lobule Segmentation Using Graph Cuts. MICCAI Challenge Workshop on Segmentation: Algorithms, Theory and Applications (2013)

166. Albin, R.L., Young, A.B., Penney, J.B.: The functional anatomy of basal ganglia disorders. Trends in neurosciences 12, 366-375 (1989)

167. Yin, H.H., Knowlton, B.J.: The role of the basal ganglia in habit formation. Nature Reviews Neuroscience 7, 464-476 (2006)

168. Brown, P.: Oscillatory nature of human basal ganglia activity: relationship to the pathophysiology of Parkinson's disease. Movement Disorders 18, 357-363 (2003)

169. Sudhyadhom, A., Haq, I.U., Foote, K.D., Okun, M.S., Bova, F.J.: A high resolution and high contrast MRI for differentiation of subcortical structures for DBS targeting: the Fast Gray Matter Acquisition T1 Inversion Recovery (FGATIR). Neuroimage 47, T44-T52 (2009)

170. Addington, J., Cadenhead, K.S., Cornblatt, B.A., Mathalon, D.H., McGlashan, T.H., Perkins, D.O., Seidman, L.J., Tsuang, M.T., Walker, E.F., Woods, S.W.: North American prodrome longitudinal study (NAPLS 2): overview and recruitment. Schizophrenia research 142, 77-82 (2012)

171. D'Haese, P.-F., Pallavaram, S., Li, R., Remple, M.S., Kao, C., Neimat, J.S., Konrad, P.E., Dawant, B.M.: CranialVault and its CRAVE tools: A clinical computer assistance system for deep brain stimulation (DBS) therapy. Medical image analysis 16, 744-753 (2012)

172. Buckner, R.L., Head, D., Parker, J., Fotenos, A.F., Marcus, D., Morris, J.C., Snyder, A.Z.: A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume. Neuroimage 23, 724-738 (2004)

173. Asman, A.J., Huo, Y., Plassard, A.J., Landman, B.A.: Multi-atlas learner fusion: An efficient segmentation approach for large-scale data. Med Image Anal 26, 82-91 (2015)

174. Panda, S., Asman, A.J., Delisi, M.P., Mawn, L.A., Galloway, R.L., Landman, B.A.: Robust Optic Nerve Segmentation on Clinically Acquired CT. Proceedings of SPIE--the International Society for Optical Engineering 9034, 90341G (2014)

175. Zitova, B., Flusser, J.: Image registration methods: a survey. Image Vision Comput 21, 977-1000 (2003)

176. Gerig, G., Jomier, M., Chakos, M.: Valmet: A new validation tool for assessing and improving 3D object segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2001, pp. 516-523. Springer, (Year)

177. Aljabar, P., Heckemann, R.A., Hammers, A., Hajnal, J.V., Rueckert, D.: Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. Neuroimage 46, 726-738 (2009)

178. Efron, B., Efron, B.: The jackknife, the bootstrap and other resampling plans. SIAM (1982)

179. Booth, J.G., Butler, R.W., Hall, P.: Bootstrap methods for finite populations. Journal of the American Statistical Association 89, 1282-1289 (1994)

180. Maldjian, J.A., Laurienti, P.J., Kraft, R.A., Burdette, J.H.: An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. Neuroimage 19, 1233-1239 (2003)