Multi-Atlas Segmentation through Rater Performance Modeling: Theory and Applications

By

Andrew Joseph Asman

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

Electrical Engineering

August, 2014

Nashville, Tennessee

Approved:

Bennett A. Landman, Ph.D.

Benoit M. Dawant, Ph.D.

Aniruddha Gokhale, Ph.D.

Adam W. Anderson, Ph.D.

Megan K. Strother, M.D.

ACKNOWLEDGMENTS

My time as a graduate student has been extremely gratifying personally, academically and scientifically. First and foremost, I am permanently indebted to my wonderful wife for her patience, understanding, and support. I don't know where I would be without your constant encouragement and belief in my abilities. To Renly, the past year has been one of the most gratifying (and exhausting) years of my life. You have taught me more in one year than any class, textbook, or lecture ever could.

Without a doubt, I would not be where I am today without the love and support of my family. I am extremely blessed to have parents that allowed me to pursue my interests and provided unwavering encouragement. To my brother, Bill, thank you for being the first scientist I ever met. You question everything and are undoubtedly a man of extraordinary principle. I have tried to emulate you throughout my entire life and these traits have been critical in my development as a student, mentor and scientist.

At Vanderbilt, I am forever grateful for the leadership and friendship provided by our leader – Bennett Landman. I have become more and more aware of how lucky we are to have your incessant encouragement, insight, and knowledge at our disposal. To my lab mates – Andrew Plassard, Benjamin Yvernault, Frederick Bryan, Xue Yang, Zhoubing Xu, Swetasudha Panda, Carolyn Lauzon, Wade Allen, Alex Dagley, and Rob Harrigan – thank you for your insights and conversation. It has been a pleasure to go on this academic ride with all of you.

Finally, I am grateful to all of our colleagues (both at Vanderbilt and around the world). Through your insights, I can depart from graduate school with the hope that I have made a difference in the world – regardless of how small that difference might be.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1. Overview

Segmentation of medical images provides a critical mechanism for relating a collection of seemingly unrelated voxel intensity values to the latent, underlying anatomical structure within an image. This structure provides localizing context that is paramount to performing nearly any analysis of medical image content. For example, segmentation is pivotal for (1) defining desired regions-of-interest, (2) providing large-scale characterization of variability in anatomical structure, (3) localizing pathological abnormalities, (4) assessing disease progression and treatment efficacy, and (5) scientific inquiry into the complex relationships between biological structure and function. This plethora of potential applications has been the driving force behind decades of research into the optimal ways to perform robust and accurate segmentation.

For well over three decades, the long-held "gold standard" for highly robust segmentation has been through expert manual delineation [1-3]. Even with the more recent image-guided interactive tools [2, 4], manual labeling is extremely time and resource consuming which limits its applicability to large-scale imaging studies. In addition to being extraordinarily time and resource consuming, manual labeling is plagued by both inter- and intra-rater variability (e.g., 10-20% by volume [5-7]). This problem of variability has led to the desire to utilize multiple raters in order to come to a consensus segmentation [8, 9] which only further complicates the problems of limited of resources. More recently, approaches have been developed that attempt to "crowd-source" the manual labeling problem using internet-based collaboration [10-16]. While certainly promising, these approaches are still years away from being a viable alternative for large-scale segmentation.

On the opposite end of the spectrum, it would be ideal if fully-automated algorithms resulted in accurate and robust segmentations. Unfortunately, wild anatomical and imaging variability often force fully-automated segmentation approaches to be highly-tuned for specific applications. For example, early medical image segmentation efforts focused on the problem of brain tissue classification (i.e., white matter vs. gray matter) [17-24], in which these tissues could be separated through direct modeling of the intensity characteristics present on the images. Assuming highly accurate pre-processing (e.g., removal of non-brain tissue), these techniques were shown to be extraordinarily successful and robust for this specific problem; however, extension of this type of framework to capture, for example, more subtle sub-cortical structures in the human brain requires *a priori* spatial context and, possibly, manual intervention [9, 25-28]. As a result of this robustness problem, automated segmentation algorithms often rely upon manual initialization and/or correction [29-32].

Given the limitations of purely manual and purely automatic segmentation, in the early 1990's atlas-based segmentation methods provided a much needed middle-ground [28, 33, 34]. In atlas-based models, spatial information is transferred from an existing labeled atlas (i.e., an example segmentation) to a previously unseen context (target subject) through a deformable registration [35-39]. An extraordinary amount of work has gone into the construction of atlases in order to represent (1) unbiased average atlases [40-42], (2) age/demographic-specfic atlases [42, 43], and (3) target-specific atlases [43, 44]. While single-atlas-based methods are quite robust, they rely upon the implicit assumption that the provided atlas can be warped through a deformable registration to find absolute correspondence with the target-of-interest. As a result of this assumption, however, the accuracy of single-atlas-based methods are inherently limited due to (1) lack of correspondence (e.g., due to morphological and pathological differences), (2) the inherent bias of the atlas image, and (3) failures in the deformable registration [19, 45-47].

Due to the restrictions of single-atlas-based methods, an alternative strategy that independently utilizes multiple atlases (i.e., multi-atlas segmentation) has come to represent the *de facto* standard baseline for robust and consistent segmentation. In multi-atlas segmentation [9, 26], multiple atlases are

separately registered to the target and the voxelwise label conflicts are resolved using label fusion [8, 11, 26, 48-62]. While pairwise registration followed by label fusion represents a typical framework for performing multi-atlas segmentation, several advancements to the multi-atlas framework have been considered, such as (1) atlas selection [53, 58, 63-66] and (2) post-processing refinement [46, 67-71]. Regardless of the intricacies of the approach, however, multi-atlas segmentation has proven to be extraordinary fruitful across a plethora of potential applications (e.g., whole-brain [26, 46, 48, 49, 51, 52, 59, 63, 65, 67, 71], hippocampus [53, 56, 61], head and neck [51, 52, 72], cardiac [57, 73, 74], prostate [58], and abdomen [75, 76]).

In the end, the fundamental question governing multi-atlas segmentation is one that has been the underpinning of decades of machine learning research: *Can we model the world though examples?* While the initial work in this field has been promising, several problems persist that limit the accuracy and prevent the widescale adoption of this framework to new problem spaces. First, while weighted voting fusion strategies [48, 56-59, 61] have come to represent the baseline fusion techniques, they remain primarily *ad hoc* and fail to provide a theoretically consistent model of multi-atlas segmentation behavior. Second, while traditional statistical fusion strategies [8, 9] provide elegant theoretical models of human labeling observation behavior, they fail to be applicable to a multi-atlas context as they (1) ignore task difficulty (i.e., some voxels are harder to label than others) [50, 77], (2) provide global estimates of rater behavior (i.e., they do not allow spatially varying performance) [49, 52, 54], (3) make implicit assumptions about highly accurate atlas-target correspondence [51, 78], and (4) neglect the hierarchical relationships between the labels exhibited in the anatomy [79].

In this dissertation, we address these challenges through a cohesive view of multi-atlas segmentation based upon rater performance modeling. For the rest of this chapter, we delve into the details surrounding multi-atlas segmentation, and outline where our work falls in the overarching structure. Finally, this chapter concludes by clearly defining the contributions provided by this dissertation.

2. Atlases

Atlases represent a fundamental unit for understanding structure and anatomical context across a population. The origins of modern research into the construction of atlases stems from the seminal work by Jean Talairach in 1988 [80] in which he developed a standardized 3D coordinate space for which brain anatomy could be analyzed. With the development of magnetic resonance imaging (MRI), the space developed by Talairach was extended to define a stereotaxic imaging space [81-85] and later utilized to make statistical inferences across a population of subjects [86-88]. Later, due to the limitations of the initial Talairach formulation (e.g., the fact that it was established using a single, post-mortem 60-year old female brain), alternative techniques for atlas construction included (1) using large numbers of subjects (e.g., the MNI-305 atlas [86]), and (2) construction of a high SNR single-subject atlas using multiple scans over time (e.g., the MNI-Colin27 atlas [89]). Finally, more recently, efforts have gone into the construction of (1) unbiased average atlases [40-42], and (2) age/demographic-specfic atlases [42, 43]. See Figure I.1 for a graphical representation of the evolution of atlases.

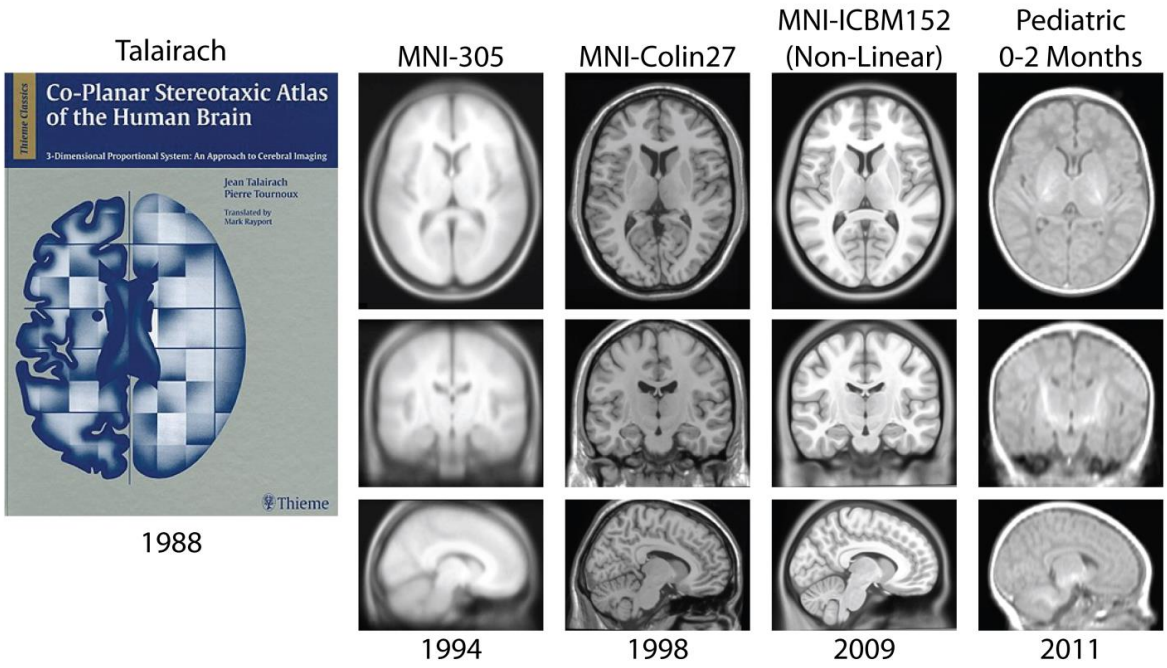


Figure I.1. The evolution of atlases. In 1988, Jean Talairach proposed a stereotaxic model for the human brain population. Since, alternative techniques for atlas construction included (1) using large numbers of subjects, (2) using multiple scans of a single subject over time, (3) unbiased average atlases, and (4) age/demographic-specific atlases.

3. Atlas-Based Segmentation

The above description of atlases is based upon the viewpoint of analyzing images within a common, consistent coordinate space (i.e., stereotaxy). As a result, subject images are generally aligned with the stereotaxic space and inferences are made about the coordinate system itself. To contrast, an alternative viewpoint of atlases is to use the inverse perspective and align the atlas space to the acquired target coordinate system. While this loses generality in terms of making inferences about populations, it gains the ability to make target-specific inferences [43, 44]. Additionally, when the atlases have associated labels to represent the underlying structure, these labels can be propagated using the atlas-target deformation field in order to infer structural information about the target subject [28] (see Section 4 for further details on image registration for atlas-based segmentation). This label propagation technique (traditionally known as atlas-based segmentation) enables an extremely straightforward technique for estimating anatomical structure. Unfortunately, atlas-based segmentation is typically relegated as a pre-processing step to many of the more popular automated segmentation algorithms [17, 19, 25, 90, 91] due to the inherent problems of using a single atlas to represent the target population (e.g., dependence on highly accurate deformable registration, and the implicit assumption that the atlas is representative of the target anatomy).

4. Registration in Atlas-Based Segmentation

The fundamental technique that enables atlas-based segmentation is image registration (see [92] for a survey of medical image registration techniques). In its essence, image registration is an optimization problem in which two images are aligned by trying to minimize some cost criterion (e.g., mean squared difference, correlation coefficient, mutual information [93]). In general, medical images are registered using a multi-stage framework in which global alignment is maximized initially, followed by a local (or voxelwise) alignment. For 3-D medical image volumes, the *global* alignment phase typically involves maximizing a rigid transformation (i.e., a 6 degree-of-freedom transformation model) [94-97] and/or an affine transformation (i.e., a 12 degree-of-freedom transformation model) [98-101]. For the

local alignment phase, the previously estimated global transformation is used to initialize a deformable non-rigid registration to maximize the voxelwise correspondence between the images [36-38, 102]. The problem of deformable non-rigid registration has been the focus of decades worth of medical imaging research. As a result, a complete review of the available techniques is outside the scope of this dissertation; however, in a recent study of 14 non-rigid registration algorithms, Klein et al [39] demonstrated the importance of registration accuracy on atlas-based segmentation techniques where it was shown that the symmetric normalization (SyN) algorithm was the most consistent performer in terms of the resulting segmentation accuracy [36].

As previously discussed, in atlas-based segmentation, the goal is to estimate the segmentation for some target image from a provided atlas image with corresponding labels via image registration. Specifically, the atlas image is registered to the target image using the previously described framework. Finally, the atlas labels can be transformed to the target coordinate system by propagating (or transferring) the atlas labels through the previously estimated transformations/deformations and the final target segmentation is estimated via an appropriate label-preserving interpolation scheme.

5. Multi-Atlas Segmentation

While using a single atlas to represent a target population is problematic, a natural and simple way to account for imaging and anatomical variability across subjects is to independently use multiple manually labeled atlases. As a result, individual failures in registration or in the applicability of an individual atlas can be overcome by aggregating the information provided by the set of atlases. This process (known as multi-atlas segmentation) was originally proposed in 2004 by Torsten Rohlfing [9] and then popularized by Rolf Heckemann on neurological data in 2008 [26]. Since its inception, multi-atlas segmentation has exploded in popularity and has been used across an extraordinary range of potential applications and imaging sequences/modalities – including, but not limited to, whole-brain [26, 46, 48, 49, 51, 52, 59, 63, 65, 67, 71], hippocampus [53, 56, 61], head and neck [51, 52, 72], cardiac [57, 73, 74], prostate [58], and abdomen [75, 76].

5.1. Typical Multi-Atlas Segmentation Workflow

The general formulation of multi-atlas segmentation, despite being quite straightforward, consists of many components (see Figure I.2 and the subsequent sections for details on each of the individual components). Put simply, the goal is to use a collection of example atlases (consisting of both intensity and labeled images) and generalize this information to a new, previously unseen context (i.e., the target image). However, as each of the atlases and the target are within in their own defined space (generally defined by the scanner acquisition), the first step is to deform the atlas images to the target coordinate system. There are many ways to accomplish this transformation, the most popular being to use a pairwise registration procedure in which each of atlas images are separately deformed to match the target image. Then, using the acquired transformations (e.g., an affine transformation or non-rigid deformation), the associated atlas labels can then be transferred (or "propagated") to the target coordinate system using an appropriate, label-preserving, interpolation scheme. Thus, after all of the atlases have been spatially

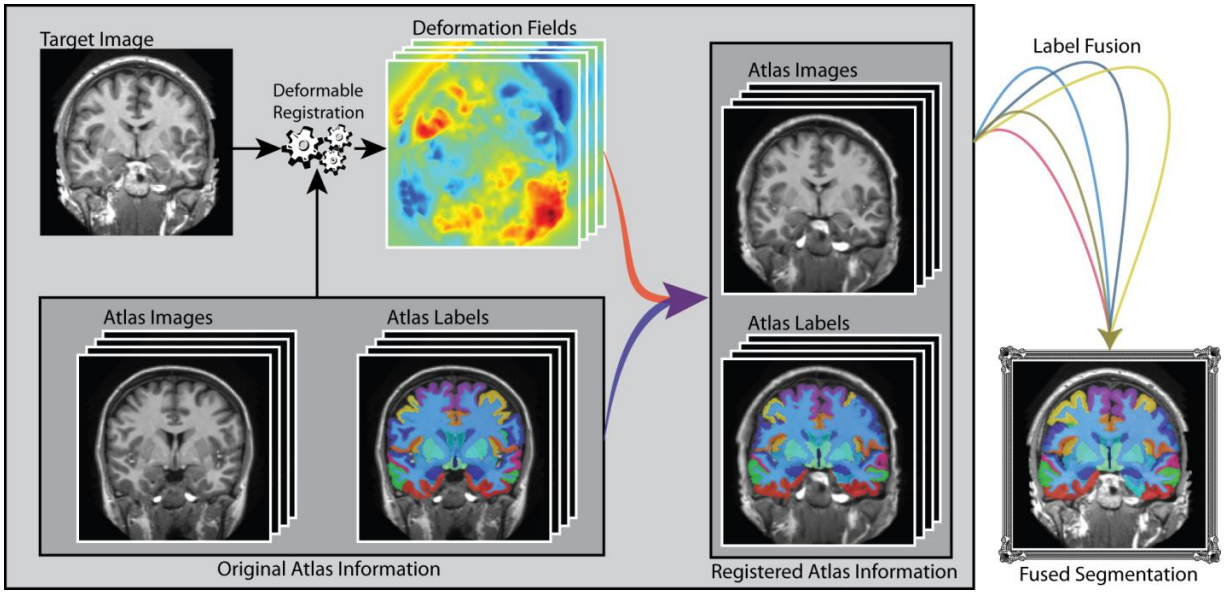


Figure I.2. A typical multi-atlas segmentation workflow. First, the target image and the atlas information are passed to a deformable registration in order to achieve a deformation field that maps the atlases to the target coordinate system. Second, the atlas information is then passed through the resulting deformation fields in order to construct the registered atlas information. Finally, all of the resulting information is then combined into a label fusion framework in order to achieve the final, fused segmentation.

normalized with the target image, we are left with a collection of label observations governing the underlying target segmentation. The voxelwise label conflicts between these observations can then be resolved using a process known as "label fusion" in order to provide a single, probabilistic estimate of the underlying segmentation. In general, there are two primary ways to perform label fusion: (1) voting-based fusion in which each of the atlases are weighted based upon some *a priori* simmilarity metric with the target image, and (2) statistical fusion in which the registered atlas (or "rater") performance is simultaneously estimated with the underlying segmentation (see Section 5 for an outline of label fusion techniques). While the above description outlines a typical multi-atlas workflow, several additional steps have been considered (a collection of which are summarized in Section 6).

6. Registration in Multi-Atlas Segmentation

Finding appropriate and optimal deformations between the atlases and the target image plays a critical role in defining multi-atlas segmentation accuracy. Obviously, if the registration fails (e.g., converges to a suboptimal local optimum as a result of highly variable anatomy, fields-of-view, or image quality), then the accuracy of the resulting propagated labels are inherently limited. Thus, the manner in which atlas information is transferred to the target context is essential in multi-atlas segmentation. Below, we consider the two most common ways in which registration is performed in multi-atlas segmentation.

6.1. Pairwise Registration

The first and most popular manner in which to propagate atlas information to the target image is through a pairwise registration procedure [9, 26, 46, 48, 49, 51-54, 57-59, 61, 63, 68]. In a pairwise registration procedure, the propagation of the atlas information to the target image is treated independently for each of the individual atlases. Thus, for each individual atlas, the previously described atlas-based registration procedure is repeated for each atlas (i.e., *globally* initialized using a rigid [94-97] or affine registration [98-101] followed by a *local* alignment using a highly deformable non-rigid registration [36-38, 102]).

The above pairwise registration procedure has been shown to be extremely successful, particularly for neurological applications. However, when extended outside of the cranial vault, (e.g., to the abdomen [75], or head and neck structures [52, 72]) more complex multi-scale, multi-level pairwise registration procedures are often implemented. For example, these type of techniques might include an initial rigid/affine segmentation that attempts to align the boney structures in the image which can often be approximated using a soft thresholding technique [72, 103, 104].

6.2. Groupwise Registration

The primary problem with a pairwise non-rigid registration framework is the excessive run time. For example, if a single-atlas registration procedure takes, on average, time T seconds, then the multi-atlas registration in a pairwise framework would inherently take approximately $R \times T$ seconds, where R is the number of registered atlases. For situations where R is very large, this amount of time can be unacceptably long and limit the applicability of multi-atlas segmentation in many clinically useful scenarios. As a result of these time limitations, there has been increased interest in performing groupwise registration procedures. Groupwise registration builds on the theoretical developments by Cootes et al [105, 106] in which shape and appearance models were constructed in order to summarize the variability in a collection of images. As a result, typical groupwise registration techniques summarize the training atlases into a common coordinate space that provides an unbiased representation of the data as a whole [107, 108]. Note that this unbiased representation are not typically an "average map" taken from the atlas images. Instead, the goal is typically to find the optimal representation in which the amount of deformation between the atlases and the obtained unbiased image is arbitrarily small. Regardless, using this type of model, all of the atlases can be spatially normalized into a common coordinate system without requiring knowledge of the target image. Then, due to the fact that the atlases are all pre-aligned, a single deformation from the group mean to the target image can be obtained and all of the atlases can be propagated through the acquired groupwise deformation.

Groupwise registration has become increasingly popular in the multi-atlas segmentation literature [44, 63, 109]. In particular, applications for (1) determining the most similar atlases for performing label fusion [63, 65, 66], (2) decreasing the computational burden of multi-atlas segmentation [109], and (3) maintaining groupwise consistent segmentations [110, 111]. While the traditional pairwise registration framework is still considered to be the gold standard for performing robust multi-atlas segmentation, these recent efforts have continually shown that the gap in performance between pairwise and groupwise registration frameworks is diminishing.

7. Label Fusion

Label fusion represents the primary focus of this dissertation. Given the target image and the registered atlas information (consisting of both intensities and labels) label fusion attempts to estimate the underlying segmentation. The concept of label fusion arose in the context of statistical machine learning. Kearns and Valiant suggested that a collection of "weak learners" (i.e., raters that are only slightly better than chance) could be fused (or "boosted") to form a "strong learner" (i.e., a single rater with arbitrarily high accuracy) [112]. This proposal was first proven a year later [113], and, with the introduction of AdaBoost [114] in 1995, the process of "boosting" multiple classifiers became widely practical and popular. Perhaps suprisingly, it wasn't until approximately 2004 that the boosting literature was successful translated to the medical imaging analysis context for the fusion of image labels [8, 9]. Today, label fusion is an extremely popular topic for ongoing research. Below, we delve into the details of label fusion and discuss the different perspectives on the optimal ways to fuse image labels.

7.1. Problem Definition

Consider a target gray-level image represented as a vector, $\boldsymbol{I} \in \mathbb{R}^{N\times 1}$. Let $\boldsymbol{T} \in \boldsymbol{L}^{N\times 1}$ be the latent representation of the true target segmentation, where $\boldsymbol{L} = \{0, \ldots, L-1\}$ is the set of possible labels that can be assigned to a given voxel. Additionally, consider a collection of R registered atlases with associated intensity values, $\boldsymbol{A} \in \mathbb{R}^{N\times R}$, and label decisions, $\boldsymbol{D} \in \boldsymbol{L}^{N\times R}$. Throughout, the index variables i,

i^* and i' will be used to iterate over the voxels, s, l, and s' over the labels, and j over the registered atlases. Given this information, the goal of any label fusion framework is to accurately estimate the following probability density function:

$$W_{si} \equiv f(T_i = s|\boldsymbol{I}, \boldsymbol{A}, \boldsymbol{D}) \tag{1.1}$$

where W_{si} can be directly interpreted as the probability that the true label at voxel i is equal to label s given the provided contextual information.

Using a standard Bayesian expansion, Eq. 1 can be re-written as

$$W_{si} = \frac{f(T_i = s)f(\boldsymbol{D}, \boldsymbol{A}|T_i = s, I)}{\sum_l f(T_i = l)f(\boldsymbol{D}, \boldsymbol{A}|T_i = l, I)} \tag{1.2}$$

Where, $f(T_i = s)$ represents the *a priori* distribution governing the underlying segmentation and $f(\boldsymbol{D}, \boldsymbol{A}|T_i = s, I)$ represents distribution governing the relationships between the observed atlas information and the latent target segmentation. Lastly, one of the most common assumptions in the label fusion literature [59] is that the observed atlas labels and the observed atlas intensities are conditionally independent resulting in

$$W_{si} = \frac{f(T_i = s)f(\boldsymbol{D}|T_i = s)f(\boldsymbol{A}|I)}{\sum_l f(T_i = l)f(\boldsymbol{D}|T_i = l)f(\boldsymbol{A}|I)} \tag{1.3}$$

while this might seem like it neglects the complex relationships between labels and intensity, the common assumption is that the information gained by direct incorporation of the target/atlas intensity relationships accurately approximates these complex relationships through the assumed conditional independence. With that said, further investigation into estimating the target label probabilities, using joint models of atlas performance is an active area of continuing research [60, 61].

Using this general framework, there are two primary fields-of-thought within the label fusion community. First, voting label fusion attempts to find optimal weights in order to determine which atlases are optimally representative in terms of some local/semi-local/global metric. Nevertheless, these techniques are primarily *ad hoc* and lack a consistent theoretical underpinning. In stark contrast, statistical fusion techniques attempt to model atlas performance using a statistically driven rater performance model. Significantly more detail on these two approaches is provided in the following two sub-sections.

7.2. Voting Label Fusion

The first, and simplest way to perform label fusion is to utilize a voting-based framework (see Figure I.3 for a flowchart demonstarting the typical workflow). In voting label fusion there are generally two primary assumptions (1) the *a priori* distribution is unnecessary as all of the registered atalses provide accurate and consistent spatial information (i.e., $f(T_i = s) = \frac{1}{L}\ \forall i \forall s$) and (2) the model probability density functions can be approximated using Parzen's window density approach [115]. As a result, a general voting-based fusion approach would simplify to

$$W_{si} = \frac{\sum_j f(D_{ij}|T_i = s) f(A_j|I)}{\sum_l \sum_j f(D_{ij}|T_i = l) f(A_j|I)} \tag{1.4}$$

Perhaps surprisingly, a majority vote, the simplest voting-based fusion strategy, has been consistently shown to result in highly robust and accurate segmentations [9, 26, 63]. In a majority vote, all atlases are weighted equally and result in Eq. 4 simplifies to

$$W_{si} = \frac{\sum_j \delta(D_{ij}, s)}{\sum_l \sum_j \delta(D_{ij}, l)} \tag{1.5}$$

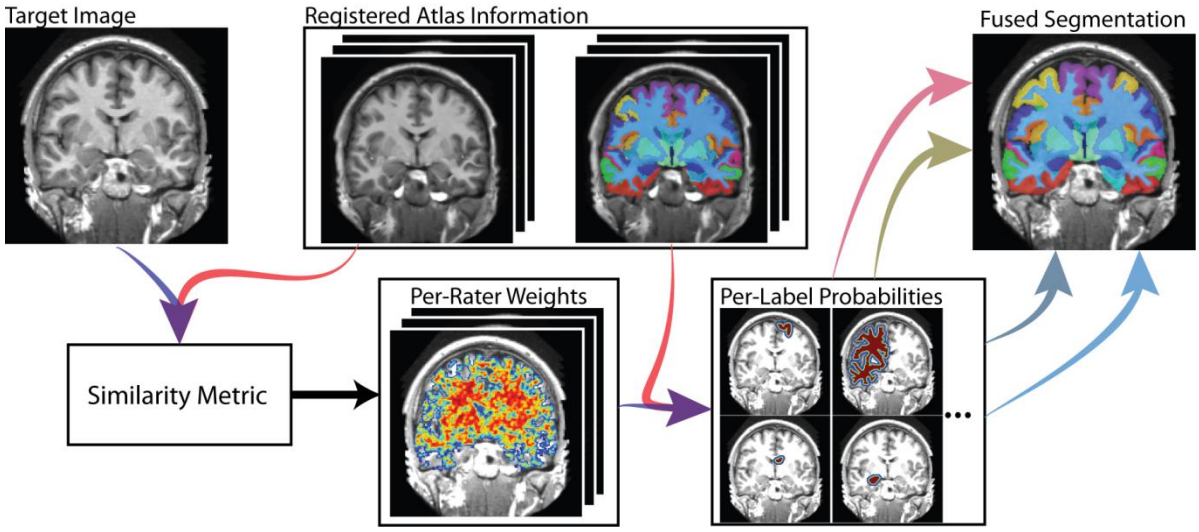


Figure I.3. A flowchart generalizing the process of voting-based label fusion. The target image and the registered atlas images are compared using a pre-defined similarity metric. The results of this comparison lead to a voxelwise weighting for each of the registered atlases. These weights are then combined with the observed labels in order to construct the probability of each label at all voxels. Finally, the fused segmentation is constructed by taking the maximum likelihood label at each voxel.

where $\delta(\cdot,\cdot)$ is the kronecker detla function.

More recently weighted voting strategiest that use global [48, 72], local [57, 59, 60], semi-local [59, 61], and non-local [56] intensity similarity metrics have demonstrated consistent improvement in segmentation accuracy. Regardless of the type of weight, however, weighted voting strategies can typically be simplified to be of the general form

$$\begin{aligned} W_{si} &= \frac{\sum_j \delta(D_{ij}, s) f(A_j|I)}{\sum_l \sum_j \delta(D_{ij}, l) f(A_j|I)} \\ &= \frac{\sum_j \omega_{ij} \delta(D_{ij}, s)}{\sum_l \sum_j \omega_{ij} \delta(D_{ij}, l)} \end{aligned} \tag{1.6}$$

where ω_{ij} is simply a weighting function governing the likelihood that the correct answer at voxel i is obtained from atlas j. Note that there are many ways to estimate these weighting functions (e.g., Gaussian intensity differences [57, 59, 61], correlation coefficients [48, 53], or mutual information [48]). For example, one of the more common weighting schemes is to use a Gaussian distribution to model the intensity differences between the target and the atlas images. Particularly for high resolution MR images of the brain, using a strict intensity difference framework has been shown to be more accurate than many of the alternative weighting schemes [48]. Using a Gaussian distribution governing the intensity difference the weighting scheme could be constructed as

$$\omega_{ij} = \frac{1}{Z_\omega} \exp\left(-\frac{(A_{ij} - I_i)^2}{2\sigma^2} \right) \tag{1.7}$$

where σ is a standard deviation parameter for the Gaussian distribution. Nevertheless, regardless of the weighting metric, the general weighted voting fusion framework remains essentially identical. With that said, additional investigation into the optimal weighting metric for a given applicationremains an open problem.

7.3. Statistical Label Fusion

In stark contrast to *ad hoc* voting, statistical fusion strategies (e.g., Simultaneous Truth and Performance Level Estimation, STAPLE [8]) directly integrate a stochastic model of rater behavior into

the estimation process (see Figure I.4 for a flowchart describing the general estimation process). Mathematically, this model of rater behavior manifests itself by augmenting the the distribgution in Eq. 1

$$W_{si} \equiv f(T_i = s|\boldsymbol{I}, \boldsymbol{A}, \boldsymbol{D}, \boldsymbol{\theta}) \tag{1.8}$$

where $\boldsymbol{\theta} \in \mathbb{R}^{R\times L\times L}$ parameterize the performance level of raters (registered atlases). Each element of $\boldsymbol{\theta}$, $\theta_{js's}$, represents the probability that rater j observes label s' given that the true label is s at a given target voxel— i.e., $\theta_{js's} \equiv f\left(D_{ij} = s'\middle|T_i = s, \boldsymbol{\theta}_j\right)$. This type of formulation of rater performance is often referred to as a "confusion matrix" (i.e., each rater has a confusion matrix that governs their labeling performance).

However, unlike voting-based label fusion frameworks, statistical fusion attempts to *simultaneously* estimate both (1) the underlying segmentation probabilities and (2) the rater performance level parameters, $\boldsymbol{\theta}$. To accomplish this, statistical fusion strategies utilize an Expectation-Maximization (EM) formulation [116]. In an EM framework, the desired model parameters are iteratively estimated using two subsequent steps (referred to as the E-step and the M-step, respectively). In the E-step, the voxelwise label probabilities are estimated using the current estimate of the rater performance level

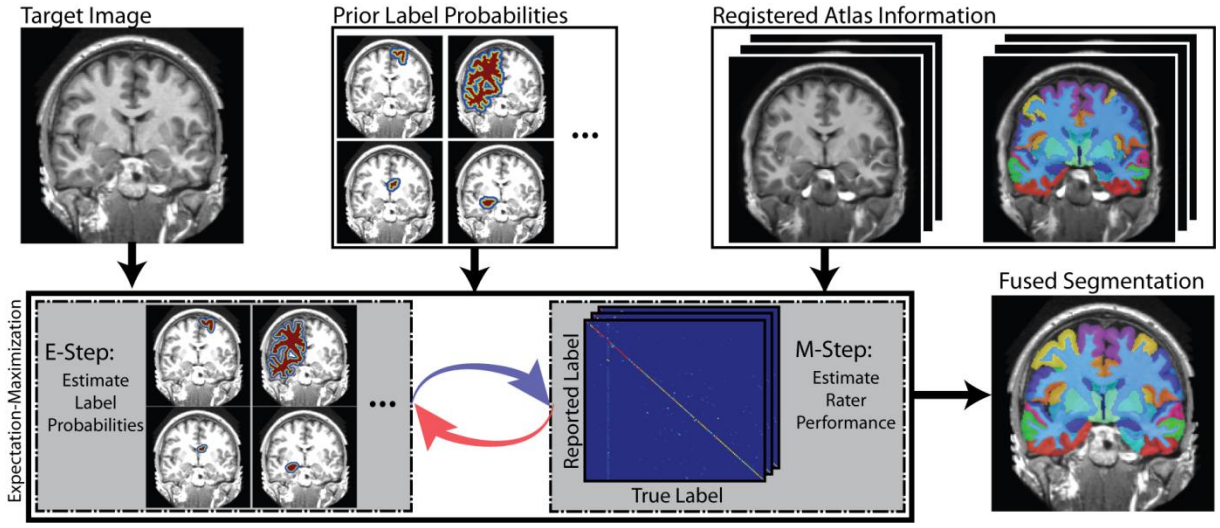


Figure I.4. A flowchart generalizing the process of statistical label fusion. Given the target image, the registered atlas information and *a priori* label probabilities, the statistical fusion process estimates the final segmentation through an Expectation-Maximization (EM) estimation process. The E- and M-steps of the EM framework are iterated until convergence of the algorithm. Lastly, given the final estimate the label probabilities, the fused segmentation is constructed by taking the maximum likelihood label at each voxel.

parameters. In the M-step, the performance level parameters are updated by finding the parameters that maximize the expected value of the conditional log-likelihood function (i.e., using the results of the previous E-step). This iterative process is then repeated until the algorithm converges (i.e., until the estimated performance level parameters cease to change beyond a pre-defined threshold).

First, for the k^{th} iteration of the E-step of the statistical fusion framework, we follow a familiar procedure and apply a Bayesian expansion to Eq. 7:

$$W_{si}^{(k)} = \frac{f(T_i = s) f(\boldsymbol{D}|T_i = s, \boldsymbol{\theta})}{\sum_l f(T_i = l) f(\boldsymbol{D}|T_i = l, \boldsymbol{\theta})} \tag{1.9}$$

Assuming conditional independence between the registered atlases (or "raters") and using a simple substitution we can then obtain

$$\begin{aligned} W_{si}^{(k)} &= \frac{f(T_i = s) \prod_j f\left(D_{ij} = s' \middle| T_i = s, \boldsymbol{\theta}\right)}{\sum_l f(T_i = l) \prod_j f\left(D_{ij} = s' \middle| T_i = l, \boldsymbol{\theta}\right)} \\ &= \frac{f(T_i = s) \prod_j \theta_{js's}}{\sum_l f(T_i = l) \prod_j \theta_{js'l}} \end{aligned} \tag{1.10}$$

where s' is simply the label observed by rater j at voxel i.

For the M-step of the algorithm, we estimate the performance level parameters by finding the parameters that maximize the expected value of the conditional log-likelihood function (i.e., using the result in Eq. 9).

$$\begin{aligned} \boldsymbol{\theta}_j^{(k+1)} &= \arg\max_{\boldsymbol{\theta}_j} \sum_i E\left[\ln f(D_{ij} = s' | T_i = s, \theta_{js's}^{(k)}) \middle| \boldsymbol{D}, \boldsymbol{\theta}^{(k)}\right] \\ &= \arg\max_{\boldsymbol{\theta}_j} \sum_i \sum_s W_{si}^{(k)} \ln f(D_{ij} = s' | T_i = s, \theta_{js's}^{(k)}) \end{aligned} \tag{1.11}$$

Noting the constraint that each row of the rater performance level parameters must sum to unity to be a valid probability mass function (i.e., $\sum_{s'} \theta_{js's}^{(k)} = 1$), we can maximize the performance level parameters for each element by using a Lagrange Multiplier (λ) [117] to formulate the constrained optimization problem. Following this procedure, we obtain

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \theta_{js's}} \left[\sum_i \sum_s W_{si}^{(k)} \ln f(D_{ij} = s' | T_i = s, \theta_{js's}^{(k)}) + \lambda \sum_{s'} \theta_{js's} \right] \\
0 &= \frac{\partial}{\partial \theta_{js's}} \left[\sum_i \sum_s W_{si}^{(k)} \ln \theta_{js's} + \lambda \sum_{s'} \theta_{js's} \right] \\
0 &= \sum_{i:D_{ij}=s} \frac{W_{si}^{(k)}}{\theta_{js's}} + \lambda \\
-\lambda &= \frac{\sum_{i:D_{ij}=s} W_{si}^{(k)}}{\theta_{js's}} \\
\theta_{js's} &= \frac{\sum_{i:D_{ij}=s} W_{si}^{(k)}}{-\lambda}.
\end{aligned}
\tag{1.12}
$$

Finally, solving for the Lagrange Multiplier leaves the final solution for each element of the performance level parameters

$$
\theta_{js's} = \frac{\sum_{i:D_{ij}=s'} W_{si}^{(k)}}{\sum_i W_{si}^{(k)}} \tag{1.13}
$$

Note that there are several considerations for (1) detection of convergence, and (2) initialization strategies for the framework. However, in the interest of brevity, we refer the reader to [8] for the intricacies of the approach.

Nevertheless, despite elegant theory and success on human raters, applications of the statistical fusion framework to the multi-atlas context have proven problematic [49, 59-61]. In response, a myriad of advancements to the statistical fusion framework have been proposed to account for (1) spatially varying task difficulty [9, 50], (2) spatially varying rater performance [49, 52, 54, 67], (3) instabilities in the rater performance level parameters [11, 55], and (4) models of hierarchical performance estimation [79]. While these advancements have been shown to dramatically improve segmentation accuracy, they still fail to incorporate useful intensity information that is critical for accurately modeling multi-atlas segmentation behavior. As a result, alternative techniques that utilize intensity information have recently been proposed that account for (1) imperfect registration correspondence [51] and (2) *ad hoc* extensions that ignore voxels based upon *a priori* similarity measures [53, 67, 118] have been considered.

To summarize, statistical fusion strategies represent a fascinating framework for estimating the underlying segmentation while simultaneously incorporating a model of registered atlas observation behavior. This theoretical consistency makes statistical fusion significantly more attractive than *ad hoc* voting-based approaches. Herein, the primary focus of this dissertation is on modifying the statistical fusion framework to more accurately model the type of label observation behavior exhibited in a multi-atlas context. In **Chapters II-V**, we propose four distinct, yet complementary, models of rater behavior that attempt to account for the limitations of the original statistical fusion formulation.

8. Additional Processing Steps

While registration (and label transfer) followed by label fusion represents the fundamental framework for performing multi-atlas segmentation, several additional processing steps have been proposed in order to increase the efficiency, robustness, and accuracy of the final segmentation. In general, these processing steps can be broken up into two distinct categories (1) pre-processing (e.g., intensity normalization, atlas selection, and patch-based approaches) and (2) post-processing (e.g., Markov Random Field regularization, intensity clustering, and learning-based wrapper methods).

8.1. Pre-Processing

8.1.1. Intensity Normalization

Given the popularity of utilizing a strict intensity difference model in order to construct local weights (see Eq. 7) for the fusion process [46, 48, 57, 59, 61], intensity normalization between images plays a critical role in segmentation accuracy. In particular, intensity normalizing MR images is critical for accurate analysis due to the fact that the intensity characteristics between MR images often varies dramatically between scanners and acquisition sequences [119-121]. Perhaps surprisingly, there is no standard technique for normalizing intensity between images for multi-atlas segmentation. The most common technique is to simply normalize the intensity between the images at a pre-defined percentiles on the images [57, 59]. Alternative techniques are to use a piece-wise linear function to map the intensities

[68] (i.e., assuming a standard gray matter, white matter, cerebro-spinal fluid intensity model). More recently, groups have considered fitting more complex functions to map the intensities between the atlas and the target image (e.g., higher order polynomials [51]). Nevertheless, due to the issues of applying these intensity normalization technique across applications, the use of alternative similarity metrics is becoming increasing popular even if it results in slightly lower accuracy for certain applications [53].

8.1.2. Atlas Selection

Atlas selection was one of the first advancements to be considered for multi-atlas segmentation. Originally proposed in 2004 [64], atlas selection is an extremely simple process in which one ignores certain atlases in the estimation process due to the fact that they are not representative of the target subject. Mathematically, atlas selection is extremely simple. Given an indicator variable, $\boldsymbol{\kappa} \in \{0,1\}^{N\times R}$, atlas selection can be applied to a majority vote through a simple manipulation of Eq.5

$$W_{si} = \frac{\sum_j \kappa_{ij}\delta\left(D_{ij}, s\right)}{\sum_l \sum_j \kappa_{ij}\delta\left(D_{ij}, l\right)} \tag{1.14}$$

where the only difference is that κ_{ij} indicates whether or not atlas j is used in determining the final segmentation at voxel i. Note, the general formulation presented in Eq. 14, enables voxelwise atlas selection. This can be simplified to perform global atlas selection by enforcing the constraint that $\kappa_{ij} = \kappa_{i'j}\ \forall i \neq i'$.

Since its inception atlas selection has been shown to result in minor, yet, statistically significant improvement over standard fusion techniques [63]. More recently, atlas selection has been extended to be part of an iterative estimation process [58], and integrated into the statistical fusion estimation process [118] which has been shown to provide additional improvement in segmentation accuracy.

It should be noted, however, that atlas selection is really just a special case of a traditional weighted voting fusion framework (i.e., a weighted voting framework where the weights are limited to a pre-defined set of binary values -- $\omega_{ij} \in \{0, 1\}$). Thus, the popularity of *explicit* atlas selection has waned due to the more intuitive and theoretically optimal models in which *implicit* and *partial (*i.e., $\omega_{ij} \in [0, 1]$)

atlas selection is possible. Additionally, it is important to note that the rater performance models implemented as part of a statistical fusion framework are a type of *implicit* atlas selection as the quality of the raters are directly modeled within the estimation framework.

8.1.3. Patch-based Approaches

While the deformable atlas-target registrations are often highly accurate on a global level, the voxelwise correspondence is often hindered by smoothness constraints and dramatic morphological disparities. As a result, many approaches have considered using a patch-based approach to reformulate the atlas label observations to more accurately match the target intensity characteristics [51, 56, 61, 74, 78]. These patch-based approaches build on the approach described as *non-local means*, a framework that emerged in the context of image de-noising [122-127]. In the non-local means framework, images are deconstructed into a collection of small volumetric patches and the similarity or correspondence between these patches is quantified to learn contextual information about the underlying image structure [122]. Specifically, in the multi-atlas context, the similarity of the atlas patches with the current target patch can be used to transform the initial labeled observations into probabilistic observations and relax the any assumptions about perfect voxelwise correspondence from the initial atlas-target registrations.

8.2. Post-Processing

Several techniques have been considered for performing a post-processing refinement of the estimated segmentation in order to improve the segmentation accuracy. While many errors in a multi-atlas segmentation context can be corrected through more efficient modeling of the atlas observation behavior, there exists a different type of error in which *consistent* mistakes are to be expected due to (1) inconsistencies in the expected labels, and (2) pathological/morphological differences that are not present on the atlases. As a result, post-processing refinement techniques have been considered that attempt to account for these types of errors.

Building upon the random field theory in imaging statistics [128], Markov Random Fields (MRFs) provide a mechanism for enforcing spatial consistency across images. In general, the idea behind

MRF integration is to regularize the *a priori* distribution by simultaneously taking into account the previous estimate of the segmentation and the relationships between the neighboring voxels. Through careful design of the neighborhood (or *clique)* structure, MRFs provide a theoretically sound procedure for incorporating the complex interactions between voxels in an image. Particularly, for early intensity-based segmentation algorithms, MRFs played in critical role in increasing the accuracy and consistency of the underlying segmentations [129, 130]. For multi-atlas segmentation, spatial consistency is typically maintained due to the smoothness constraints of typical registration algorithms. As a result, although MRFs have been shown to provide accuracy improvements for multi-atlas segmentation [8], the observed improvement is typically minimal.

Another type of post-processing refinement is to perform a meta-analysis framework in order to enforce desired constraints (i.e., consistency with the training data). For instance, the idea of applying the intensity-driven EM segmentation after a multi-atlas segmentation is becoming increasingly popular [68, 69]. However, this type of formulation is limited to scenarios where the intensity characteristics between the desired anatomical structures can be probabilistically separated. For example, in multi-organ abdomen segmentation, many of the desired organs have identical, or nearly identical, intensity characteristics on CT. As a result, EM refinement techniques are often unable to discover distributions that uniquely identify the individual organs, and the approach would rely on extremely accurate *a priori* structural context.

Lastly, machine learning techniques that construct classifiers (e.g., via AdaBoost) for correcting consistent segmentation errors are becoming increasingly popular in the segmentation literature (e.g., [70, 76]). These type of "wrapper" methods have been shown to provide consistent improvement in segmentation accuracy, however, determining optimal techniques for feature selection criteria and initializing model parameters represent fascinating areas of continuing research..

9. Contributions

The primary contributions of this dissertation are as follows. In **Part 1**, we demonstrate advancements to the statistical fusion framework that (1) provide a theoretical basis for estimating labeling task difficulty (**Chapter II**), (2) demonstrate a direct mechanism for characterizing spatially varying performance (**Chapter III**), (3) seamlessly incorporate intensity information into the statistical fusion framework via a reformulation from a non-local means perspective (**Chapter IV**), and (4) estimate a general model of hierarchical performance (**Chapter V**). Next, in **Part 2**, the benefits of these theoretical advancements are illustrated for: (1) detection of imaging abnormalities and anomalies (**Chapter VI**), (2) segmenting the spinal cord's internal structure through structural shape and appearance modeling (**Chapter VII**), and (3) removing the need for computationally expensive deformable registration in whole-brain multi-atlas segmentation via machine learning mechanisms (**Chapter VIII**). Finally, we conclude by summarizing the contributions and addressing avenues for further extension and exploration (**Chapter IX**). Specifically we:

1. We extend the statistical fusion framework in order to **account for spatially-varying task difficulty**. While the traditional statistical fusion framework assumes that all voxels represent the same underlying difficulty, this has been consistently shown to be inconsistent with observed rater performance models. We augment the traditional statistical fusion framework with consensus levels. Through this augmentation, we simultaneously estimate the likelihood that each voxel belongs to each of the available consensus levels. This algorithm – Consensus Level, Labeler Accuracy, and Truth Estimation (COLLATE), (i) provides a theoretical motivation for ignoring consensus voxels, and (ii) provides statistically significant improvement over traditional statistical fusion techniques.
2. We extend the statistical fusion framework to **allow for spatially-varying performance**. Particularly in a multi-atlas segmentation context, it is to be expected that a given atlas would exhibit varying quality depending upon the quality of the registration. While traditional techniques assume a single global representation of rater performance, we provide a simple

mechanism for estimating a smooth, voxelwise estimate of rater performance in order to account for these inconsistencies in rater performance. This algorithm – Spatial STAPLE, represents the first statistical fusion algorithm that enables local characterization of rater performance.

3. We **derive a reformulation of the traditional statistical fusion framework from a non-local means perspective** to account for inconsistencies in the atlas-target registrations. This algorithm – Non-Local STAPLE, represents the first statistical fusion algorithm that (i) creates a cohesive theoretical model specifically targeting registered atlas observation behavior, and (ii) seamlessly incorporates intensity into the core of the STAPLE estimation framework. As a result, NLS largely overcomes the need for high-quality non-rigid registration and large numbers of atlases.
4. We propose a novel statistical fusion framework to **estimate a generalized model of hierarchical performance**. Given an *a priori* model of the hierarchical label relationships for a given segmentation task, the proposed model provides a straightforward mechanism for simultaneously estimating multiple (hierarchical) confusion matrices for each rater and is highly amenable to many of the state-of-the-art advancements to the statistical fusion framework.
5. Typically, multi-atlas segmentation is limited to "in-atlas" applications (e.g., applications where the atlases are anatomically and structurally indicative of the target image). We propose **a technique to estimate the out-of-atlas (OOA) likelihood for every voxel in the target image**. The OOA approach provides an intuitive and fully general abnormality/outlier detection framework that (i) uses multiple normal atlases to limit the inherent bias of using a single atlas and avoid the need for non-rigid registration, and (ii) can be used in a large number of potential applications.
6. We propose the **first approach for fully automated segmentation of cervical spinal cord internal structure using a groupwise slice-based multi-atlas registration framework**.

Specifically, we provide a method for (i) pre-aligning the slice-based atlas information into a common, groupwise-consistent coordinate system, (ii) constructing a model describing spinal cord variability (i.e., "eigenspines"), (iii) registering the target image slice to the model space using a simultaneous intensity- and model-driven cost function, and (iv) estimating a final segmentation by fusing the provided atlas information.

7. We propose **geodesic learner fusion (GLF), a framework for rapidly and accurately replicating the highly accurate, yet computationally expensive, multi-atlas segmentation framework based on fusing geodesically appropriate learners**. In the largest whole-brain multi-atlas study ever reported, we (i) estimate a low-dimensional representation for selecting geodesically appropriate example images, and (ii) build AdaBoost learners that map a weak initial segmentation to the multi-atlas segmentation result. Thus, to segment a new target image we simply project the image into the low-dimensional space, construct a weak initial segmentation, and fuse the trained, geodesically appropriate, learners.

10. Previous Publications

Many contributions of this dissertation have been published. Advancements to the statistical fusion framework are discussed for characterizing task difficulty [50, 77], spatially varying performance [49, 52], imperfect correspondence [51, 78], and hierarchical performance estimation [79]. The ability to detect abnormalities, pathologies, and quality control issues are shown in [131]. The groupwise multi-atlas segmentation framework for segmentation of the spinal cord's internal structure is discussed in [132, 133]. Additionally, advancements to the field of collaborative labeling provide support for many of the contributions of this dissertation [10-16].

PART 1

THEORY

The first part of this thesis focuses on theoretical advancements to the statistical label fusion framework. Building on the seminal Simultaneous Truth and Performance Level Estimation (STAPLE) [8], we present theoretical reformulations to more accurately characterize rater (or atlas) performance. Specifically, these advancements provide methods for: (1) estimating task difficulty (**Chapter II**), (2) formulating spatially varying performance (**Chapter III**), (3) accounting for registration uncertainty and imperfect correspondence (**Chapter IV**), and (4) estimating hierarchically consistent models of performance (**Chapter V**). Together, these theoretical advancements provide powerful mechanisms for more accurately understanding and estimating rater-driven models and, thus, more accurately estimating the desired target segmentations.

CHAPTER II

FORMULATING TASK DIFFICULTY

1. Introduction

The label fusion problem arose in the context of statistical machine learning. Kearns and Valiant suggested that a collection of "weak learners" (raters that are only slightly better than chance) could be fused ("boosted") to form a "strong learner" (a single rater with arbitrarily high accuracy) [112]. This proposal was first proven about a year later [113], and the process of "boosting" became widely practical and popular with the presentation of AdaBoost [114]. Statistical methods using automated results or complete data sets from several different human raters have been proposed to simultaneous estimate (1) the rater performance level and (2) the "ground truth" [8, 9, 134]. The algorithm presented by Warfield et al. provided a simultaneous estimation of both performance level parameters of expert segmentations and an estimation of the "ground truth" [8]. Extensions to this approach were introduced by Rohlfing et al [9]. These algorithms are based upon a maximum likelihood/maximum *a posteriori* approach (e.g. Simultaneous Truth and Performance Level Estimation, STAPLE [8]). When operating under the assumption that the raters performing the segmentations are collectively unbiased and independent, these algorithms increase the accuracy of a single labeling by probabilistically fusing multiple less accurate delineations. These statistical approaches have been widely used in atlas-fusion techniques [26] and have been extended to handle continuous (scalar or vector) images [55, 135, 136].

Despite the recent advancements in the field of label fusion, there exists a fundamental limitation in the way that these algorithms compute performance level parameters of the raters and, thus, the estimation of the true segmentation: the observed model of rater behavior when dealing with human raters is not particularly accurately modeled by the generative model of rater behavior used by STAPLE (and its

descendants). Intuitively (and empirically – see Figure II.1), raters tend to miss at a very small subset of the actual voxels present in a data set. These voxels tend to be boundary pixels and voxels where the value is ambiguous for one reason or another. This problem manifests itself, in many cases, by creating estimations of the rater performance parameters that are biased towards certain labels.

For example, imagine a truth model where there are only two labels present. One of the labels is the background, which composes a huge percentage of the total data set and the other is a label that is only present as a small circle in the middle of the truth model. Additionally, the only voxels where there is contention about the true label are the voxels that define the boundary between the background and the small label. If the observed model of rater behavior holds, the STAPLE estimate of rater behavior would estimate that the raters are very good at the background label and very bad at the small label, when, in actuality, the problems of the raters are directly related to their ability to delineate the boundary between the labels. Instead of the entire observation, there is a small subset of voxels that determine the quality of the raters, and the *consensus* voxels should not be as heavily weighted when determining the rater performance parameters. Herein, we present a robust statistical label fusion algorithm through *consensus level*, labeler accuracy and truth estimation (COLLATE). By simultaneously characterizing and estimating this additional *consensus*, we capture a more realistic model of rater behavior to more accurately estimate both rater performance and truth labels. The performance of COLLATE is characterized in simulation and with empirical data (i.e., labels provided by human raters).

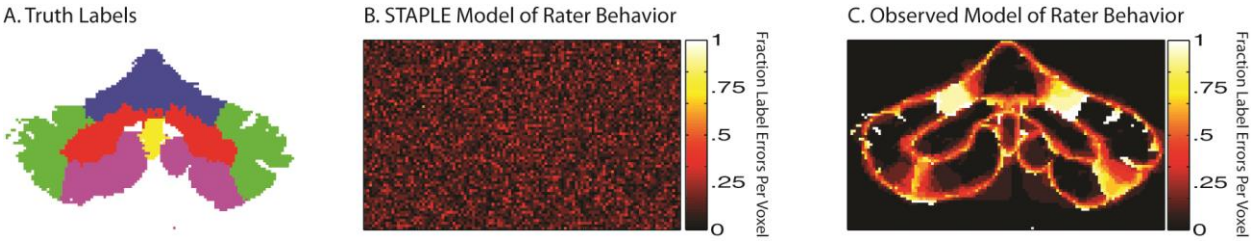


Figure II.1. The inaccuracies of the STAPLE model of rater behavior. A representative slice from the truth model is shown in (A). The expected STAPLE model of rater behavior can be seen in (B). STAPLE operates under the assumption that there is a uniform probability that any given rater would mis-label a given voxel. The observed model of rater behavior can be seen in (C). The primary difference between (B) and (C) is that the human raters showed a clear inclination to mislabel boundary pixels and other ambiguous regions.

Throughout this chapter the terms *confusion* and *consensus* will be used to characterize the likelihood that a rater makes a mistake at a given voxel. These polar terms are used as a qualitative description of the quantitative *consensus level*. For example, a voxel that is determined to have a high consensus level is considered to be a voxel where there is high *consensus* and low *confusion* (i.e. it is unlikely that a rater would make a mistake at this voxel). Alternatively, a voxel that is determined to have a low consensus level is considered to be a voxel where there is high *confusion* and low *consensus (*i.e. there is a high probability that a rater would make a mistake at this voxel).

This chapter is organized in the following manner. In Section 2, the COLLATE algorithm is described. Techniques for initializing the algorithm, detecting convergence, and the recommended method of setting the model parameters is described. In Section 3, the COLLATE algorithm is compared to traditional STAPLE on a series of experiments and simulations. One of the simulations demonstrates the sensitivity of the data-adaptive priors defined in Section 2. Additional implementations include simulations using the modified COLLATE model of rater behavior, an approximation of a human model where raters miss at boundaries, the STAPLE model of rater behavior and an empirical experiment.

2. Theory

The following derivation of the COLLATE closely follows the approach of Warfield, *et al* [8].

2.1. Problem Definition

As in the Warfield approach, consider an image of N voxels with the task of determining the correct label for each voxel in that image. Also consider a collection of R raters that provide an observed delineation for each of N voxels exactly once. Herein, the index variable *i* will be used to iterate over the N voxels and the index variable *j* will be used to iterate over the R raters. The set of labels, **L,** represents the set of possible values that a rater can assign to all N voxels. Let **D** be an N x R matrix describing the labeling decisions of all R raters at all N voxels where $D_{ij} \in \{0, 1, \dots, L-1\}$. Let **T** be a vector of N elements that represents the hidden true segmentation for all voxels, where $T_i \in \{0, 1, \dots, L-1\}$.

In addition to the traditional model, consider a vector of N elements, **C,** that represents a characterization of the *consensus* or *confusion* of each voxel at one of F level of possible consensus. All elements in this vector, $C_i \in \{0, 1, \ldots, F-1\}$, indicate whether voxel i is a voxel of *confusion* ($C_i = 0$) or a voxel of some level of *consensus* ($C_i > 0$). It is important to note that the terms *consensus* and *confusion* are polar terms that are describing the same phenomenon from opposite perspectives. As the value of C_i increases the amount of *confusion* about voxel *i* decreases, while, conversely, the amount of *consensus* about voxel *i* increases. We present theory for a multi-consensus level framework. However, only a closed form solution for the binary consensus level solution is derived. This vector will subsequently be referred to as the "consensus level vector." The E-M algorithm presented in this paper,

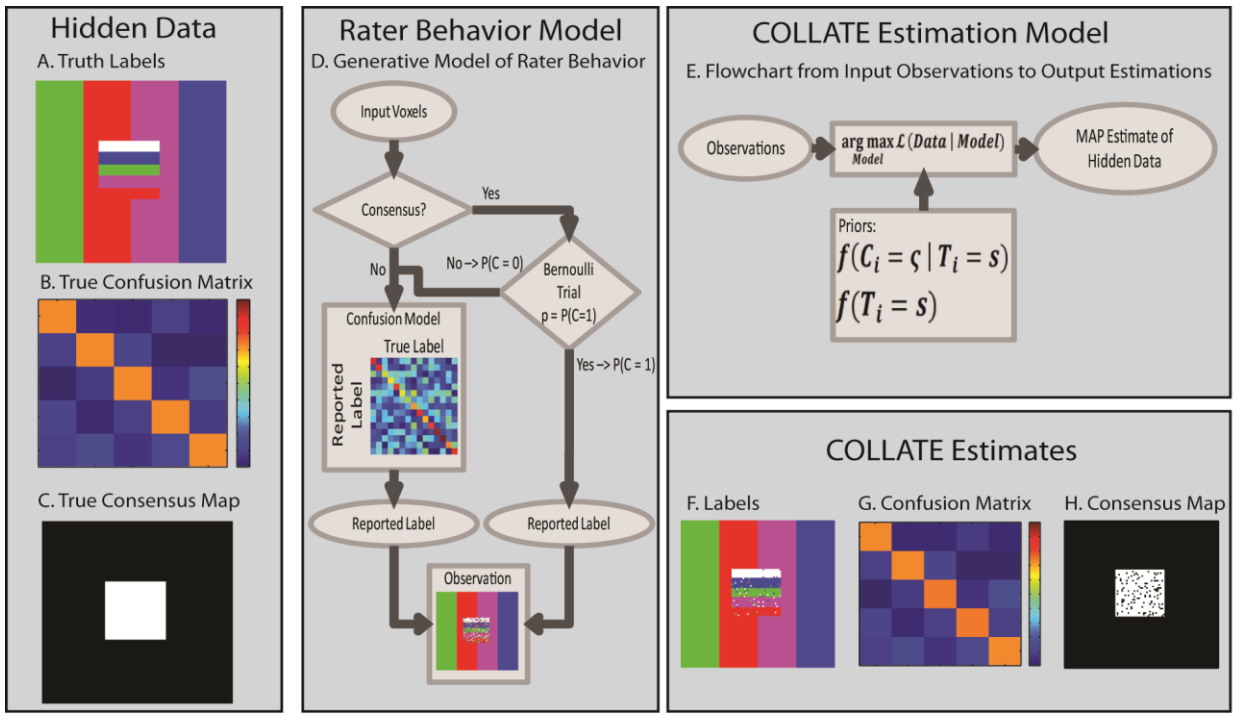


Figure II.2. The COLLATE model. The hidden data in the COLLATE E-M algorithm can be seen in (A), (B) and (C). These images (the true labels, rater confusion matrices and consensus map) represent the complete set of data that COLLATE attempts to estimate. The generative model of rater behavior can be seen in (D). This flowchart shows the path from an input voxel on some clinical data to a single observation. A flowchart demonstrating the way in which COLLATE takes input observations and estimates the hidden data can be seen in (E). Note the inclusion of priors in conditional probability that is estimated to generate the maximum *a posteriori* estimate of the hidden data. Example estimates of the hidden data can be seen in (F), (G) and (H).

will estimate the probability that voxel *i* belongs to each consensus level in the E-Step, and these estimated probabilities will be crucial in weighting each voxel when estimating the performance level parameters in the M-Step.

A characterization of the R raters' performance is characterized by $\boldsymbol{\Theta}$, where each element, Θ_j, is an L x L confusion matrix where each element in the matrix quantifies the probability that rater j will assign label s' to a voxel when the true label is s. For reference, the perfect rater would have a confusion matrix of the identity matrix. Let the complete data be (**D**, **T**, **C**) and let the probability mass function of the complete data be $f(\boldsymbol{D}, \boldsymbol{T}, \boldsymbol{C}|\boldsymbol{\theta})$.

2.2. COLLATE Algorithm

The goal of COLLATE is to accurately estimate the performance level parameters of the R raters given the rater segmentation decisions, the estimation of the truth, and the estimation of the consensus level vector (see Figure II.2). The estimated performance level parameters will be selected such that they maximize the complete data log likelihood function

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ln f(\boldsymbol{D}, \boldsymbol{T}, \boldsymbol{C}|\boldsymbol{\theta}) \tag{2.1}$$

It is assumed that the segmentation decisions are all conditionally independent given the true segmentation and the performance level parameters, that is $(D_{ij}|C_i T_i \theta_j) \perp (D_{ij'}|C_i T_i \theta_{j'}) \forall \mathrm{j} \neq \mathrm{j}'$. This model expresses the assumption that the raters derive their segmentations of the same image independently from one another and that the quality of the result of the segmentation is captured by the estimation of the performance level parameters.

Our version of the expectation-maximization (E-M) algorithm used to solve (1) is now presented. The complete data used to solve this E-M algorithm is the observed data, **D**, and the true segmentation of each voxel **T** augmented with the consensus level vector, **C**. The true segmentation **T** and the consensus level vector, **C**, are regarded as the missing or hidden data, and are unobservable. Let $\boldsymbol{\theta}_j$ be the covariance, or confusion, matrix associated with rater j and let

$$\boldsymbol{\Theta} = [\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \ldots, \boldsymbol{\Theta}_{\mathrm{R}}] \tag{2.2}$$

be the complete set of unknown parameters for the R segmentations. Let $f(\boldsymbol{D}, \boldsymbol{T}, \boldsymbol{C}|\boldsymbol{\theta})$ denote the probability mass function of the random vector corresponding to the complete data. The complete data log likelihood function is presented as

$$\ln L_c\{\boldsymbol{\theta}\} = \ln f(\boldsymbol{D}, \boldsymbol{T}, \boldsymbol{C}|\boldsymbol{\theta}). \tag{2.3}$$

The E-M algorithm approaches the problem of maximizing the incomplete data log likelihood equation

$$\ln L\{\boldsymbol{\theta}\} = \ln f(\boldsymbol{D}|\boldsymbol{\theta}) \tag{2.4}$$

by proceeding iteratively with estimation and maximization of the complete data log likelihood function. As the complete data log likelihood function is not observable, it is replaced by its conditional expectation of the observable data **D** given the current estimate of $\boldsymbol{\Theta}$. Computing the conditional expectation of the complete data log likelihood function is referred to as the E-step, and identifying the parameters that maximize this function is referred to as the M-step.

In more detail, let $\boldsymbol{\Theta}^{(0)}$ be some initial value for $\boldsymbol{\Theta}$. Then, on the first iteration, the E-step requires the calculation of

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)}) &\equiv E[\ln f(\boldsymbol{D}, \boldsymbol{T}, \boldsymbol{C}|\boldsymbol{\theta})|\boldsymbol{D}, \boldsymbol{\theta}^{(0)}] \\ &= \sum_{\boldsymbol{T}} f(\boldsymbol{D}, \boldsymbol{T}, \boldsymbol{C}|\boldsymbol{\theta}) f\left(\boldsymbol{T}, \boldsymbol{C}\middle|\boldsymbol{D}, \boldsymbol{\theta}^{(0)}\right). \end{aligned} \tag{2.5}$$

The M-step requires the maximization of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)})$ over the parameter space of $\boldsymbol{\theta}$. That is, we choose $\boldsymbol{\theta}^{(1)}$ such that

$$Q\left(\boldsymbol{\theta}^{(1)}\middle|\boldsymbol{\theta}^{(0)}\right) \geq Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(0)}\right) \tag{2.6}$$

for all $\boldsymbol{\theta}$. The E-step and the M-step are then repeated as above where at each iteration k, the current estimate $\boldsymbol{\theta}^{(k)}$, the observed data **D** are used to calculate the conditional expectation of the complete data log likelihood function, and then the estimate of $\boldsymbol{\theta}^{(k+1)}$ is found by maximizing $Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(k)}\right)$. The E- and M- steps are repeated until convergence.

The performance parameters at iteration k that maximize the conditional expectation of the log likelihood function are given by

$$\begin{aligned}\boldsymbol{\theta}^{(k)} &= \arg\max_{\boldsymbol{\theta}} E[\ln f(\boldsymbol{D},\boldsymbol{T},\boldsymbol{C}|\boldsymbol{\theta})\big|\boldsymbol{D},\boldsymbol{\theta}^{(k-1)}] \\ &= \arg\max_{\boldsymbol{\theta}} E[\ln \frac{f(\boldsymbol{D},\boldsymbol{T},\boldsymbol{C},\boldsymbol{\theta})}{f(\boldsymbol{\theta})}|\boldsymbol{D},\boldsymbol{\theta}^{(k-1)}]\end{aligned} \tag{2.7}$$

Thus, on multiplying by $\frac{f(\boldsymbol{T},\boldsymbol{C},\boldsymbol{\theta})}{f(\boldsymbol{T},\boldsymbol{C},\boldsymbol{\theta})}$

$$\boldsymbol{\theta}^{(k)} = \arg\max_{\boldsymbol{\theta}} E[\ln \frac{f(\boldsymbol{D},\boldsymbol{T},\boldsymbol{C},\boldsymbol{\theta})f(\boldsymbol{T},\boldsymbol{C},\boldsymbol{\theta})}{f(\boldsymbol{T},\boldsymbol{C},\boldsymbol{\theta})f(\boldsymbol{\theta})}|\boldsymbol{D},\boldsymbol{\theta}^{(k-1)}] \tag{2.8}$$

which yields

$$\boldsymbol{\theta}^{(k)} = \arg\max_{\boldsymbol{\theta}} E[\ln f(\boldsymbol{D}|\boldsymbol{T},\boldsymbol{C},\boldsymbol{\theta})f(\boldsymbol{T},\boldsymbol{C})|\boldsymbol{D},\boldsymbol{\theta}^{(k-1)}] \tag{2.9}$$

where $\boldsymbol{\theta}^{(k)}$is the estimate of the performance level parameters of the raters after the k[th] iteration of the algorithm. The last step operates under the assumption that **T** and **C** are independent of the performance level parameters, i.e. $f(\boldsymbol{T},\boldsymbol{C},\boldsymbol{\theta}) = f(\boldsymbol{T},\boldsymbol{C})f(\boldsymbol{\theta})$.

2.3. E-Step: Estimation of the Voxelwise Label Probabilities

In this section, the estimator for the unobserved true segmentation is derived. We first derive an expression for the conditional probability density function of the true segmentation and the consensus level vector at each voxel given the raters decisions, and the previous estimate of the performance parameters.

In order to maintain a compact representation of the result, the conditional probability of the true segmentation at each voxel is represented using a common notation for E-M algorithms.

$$\begin{aligned}W_{si\varsigma}^{(k-1)} &\equiv f(T_i = s, C_i = \varsigma|\boldsymbol{D_i},\boldsymbol{\theta}^{(k-1)}) \\ &= \frac{f(T_i = s, C_i = \varsigma)\prod_j f\left(D_{ij}\middle|T_i = s, C_i = \varsigma, \theta_j^{(k-1)}\right)}{\sum_{\varsigma'}\sum_{s'} f(T_i = s', C_i = \varsigma')\prod_j f\left(D_{ij}\middle|T_i = s', C_i = \varsigma', \theta_j^{(k-1)}\right)} \\ &= \frac{f(C_i = \varsigma \mid T_i = s)f(T_i = s)\prod_j f\left(D_{ij}\middle|T_i = s, C_i = \varsigma, \theta_j^{(k-1)}\right)}{\sum_{\varsigma'}\sum_{s'} f(C_i = \varsigma' \mid T_i = s')f(T_i = s')\prod_j f\left(D_{ij}\middle|T_i = s', C_i = \varsigma', \theta_j^{(k-1)}\right)}\end{aligned} \tag{2.10}$$

where $W_{si\varsigma}^{(k-1)}$, the weight variable, indicates the probability of the true segmentation at voxel i being equal to label s, with consensus level value ς. This representation is different from the traditional STAPLE representation of the weight variable due to the presence of the consensus level vector. For example, the value described by $W_{si0}^{(k)}$ represents the probability that voxel i is equal to label s for the k^{th} iteration and is a voxel that is likely to be confused by a given rater. The matrix constructed by considering this value at all N voxels and for all L labels is referred to later as the "consensus map." The result of augmenting the weight variable with the consensus level vector is that *consensus* voxels are isolated so that they can be weighted less heavily when computing the rater confusion matrices. This results in an unbiased estimate of rater quality where the proportion of a given label in a truth model is significantly less influential than in the STAPLE algorithm.

2.4. M-Step: Estimation of the Performance Fields via Maximization

Given the estimated weight variable $W_{si\varsigma}^{(k-1)}$, which represents the conditional probability that the true segmentation of voxel i is equal label s with consensus level value ς, it is now possible to estimate the rater performance parameters that maximize the conditional expectation of the complete data log likelihood function. Considering each rater separately, we find the parameter estimates $\theta_j^{(k)}$ by

$$\begin{aligned}\theta_j^{(k)} &= \arg\max_{\theta_j} \sum_i E[\ln f(D_{ij}|T_i, C_i, \theta_j)|\boldsymbol{D}, \theta_j^{(k-1)}] \\ &= \arg\max_{\theta_j} \sum_i \sum_s \sum_\varsigma W_{si\varsigma}^{(k-1)} \ln f(D_{ij}|T_i = s, C_i = \varsigma, \theta_j) \\ &= \arg\max_{\theta_j} \sum_{s'} \sum_{i:D_{ij}=s'} \sum_s \sum_\varsigma W_{si\varsigma}^{(k-1)} \ln f(D_{ij} = s'|T_i = s, C_i = \varsigma, \theta_j).\end{aligned} \tag{2.11}$$

We determined that

$$f(D_{ij} = s'|T_i = s, C_i = \varsigma, \theta_j) = (1 - p(\varsigma))I(s = s') + p(\varsigma)f(D_{ij} = s'|T_i = s, \theta_j) \tag{2.12}$$

where $I(s = s')$ is the indicator function. Plugging (11) into (12) yields

$$\theta_j^{(k)} = \arg\max_{\theta_j} \sum_{s'} \sum_{i:D_{ij}=s'} \sum_{s} \sum_{\varsigma} W_{si\varsigma}^{(k-1)} \ln((1-p(\varsigma))I(s=s') + p(\varsigma) f(D_{ij}=s'|T_i=s,\theta_j))$$
$$= \arg\max_{\theta_j} \sum_{s'} \sum_{i:D_{ij}=s'} \sum_{s} \sum_{\varsigma} W_{si\varsigma}^{(k-1)} \ln\left(\left(1-p(\varsigma)\right)I(s=s') + p(\varsigma)\theta_{js's}\right).$$

Note the constraint that each row of the rater parameter matrix must sum to one in order to be a probability mass function

$$\sum_{s'} \theta_{js's} = 1. \tag{2.14}$$

The rater performance parameters can be maximized through the constrained optimization problem

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta_{jn'n}} \left[\sum_{s'} \sum_{i:D_{ij}=s'} \sum_{s} \sum_{\varsigma} W_{si\varsigma}^{(k-1)} \ln\left(\left(1-p(\varsigma)\right)I(s=s') + p(\varsigma)\theta_{js's}\right) \right] \\ &= \sum_{i:D_{ij}=n'} \sum_{\varsigma} W_{ni\varsigma}^{(k-1)} \frac{p(\varsigma)}{(1-p(\varsigma))I(n=n') + p(\varsigma)\theta_{jn'n}} + \lambda \end{aligned} \tag{2.15}$$

where λ is a Lagrange multiplier.

In order to represent the solution for $\theta_{jn'n}$ there are two cases that need to be considered. First, the $n \neq n'$ case (off-diagonal) which can be shown to be equal to

$$\theta_{jn'n,n\neq n'} = \frac{\sum_{i:D_{ij}=n'} \sum_{\varsigma} W_{ni\varsigma}^{(k-1)}}{-\lambda}. \tag{2.16}$$

Up until this point, the theory presented has been for the generic multi-consensus level approach for COLLATE. However, it is more involved to analytically solve for the $n = n'$ (on-diagonal) case for $\theta_{jn'n}$. Function optimization methods (e.g. simplex, annealing, etc) could be applied to numerically solve for the case of an arbitrary number of consensus levels.

For simplicity of representation in this paper the binary case, where $\varsigma\epsilon\{0,1\}$ and $p(1) = 1 - p(0)$, is solved below

$$
\begin{aligned}
0 &= \sum_{i:D_{ij}=n'} \sum_{\varsigma} W_{ni\varsigma}^{(k-1)} \frac{p(\varsigma)}{1-p(\varsigma)+p(\varsigma)\theta_{jn'n,n=n'}} + \lambda \\
-\lambda &= \sum_{i:D_{ij}=n'} W_{ni0}^{(k-1)} \frac{p(0)}{(1-p(0))+p(0)\theta_{jn'n,n=n'}} + W_{ni1}^{(k-1)} \frac{1-p(0)}{p(0)+(1-p(0))\theta_{jn'n,n=1}}
\end{aligned}
\tag{2.17}
$$

The solution for $\theta_{jn'n,n=n'}$ was obtained using Mathematica (Wolfram Research, Champaign, IL). For ease of representation, three dummy variables (a, b, and c) are declared below to solve for $\theta_{jn'n,n=n'}$

$$a = \lambda \prod_{\varsigma} p(\varsigma) \tag{2.18}$$

$$b = \lambda \left(\sum_{\varsigma} p(\varsigma)^2 \right) + \left(\prod_{\varsigma} p(\varsigma) \right) \sum_{i:D_{ij}=n',n=n'} \sum_{\varsigma} W_{ni\varsigma}^{(k-1)} \tag{2.19}$$

$$c = \left(\sum_{i:D_{ij}=n',n=n'} \sum_{\varsigma} p(\varsigma)^2 W_{ni\varsigma}^{(k-1)} \right) + \lambda \prod_{\varsigma} \big(p(\varsigma)\big). \tag{2.20}$$

The final solution for $\theta_{jn'n,n=n'}$ is

$$\theta_{jn'n,n=n'} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \tag{2.21}$$

The remaining step is to solve for λ. This can be accomplished using the constraint defined in (16), which is redefined below due to the separation of the $n = n'$ and $n \neq n'$ cases

$$\theta_{jn'n,n=n'} + \sum_{n':n\neq n'} \theta_{jn'n,n\neq n'} = 1. \tag{2.22}$$

Given that both $\theta_{jn'n,n=n'}$ and $\theta_{jn'n,n\neq n'}$ are functions of λ it is possible, after some algebra, to represent the closed form solution for λ. As with the solution for $\theta_{jn'n,n=n'}$, three dummy variables are declared to ease the representation of the solution

$$\alpha = 1 \tag{2.23}$$

$$\beta = \sum_{n':n\neq n'} \sum_{i:D_{ij}=n'} \sum_{\varsigma} W_{ni\varsigma}^{(k-1)} + \sum_{i:D_{ij}=n',n=n'} \sum_{\varsigma} p(\varsigma) W_{ni\varsigma}^{(k-1)} \tag{2.24}$$

$$\gamma = \left(\prod_{\varsigma} p(\varsigma)\right)\left(\sum_{n':n\neq n'}\sum_{i:D_{ij}=n'}\sum_{\varsigma} W_{ni\varsigma}^{(k-1)}\right)\left(\sum_{i}\sum_{\varsigma} W_{ni\varsigma}^{(k-1)}\right) \tag{2.25}$$

The solution for λ can be written as

$$\lambda = \frac{-\beta \pm \sqrt{\beta^2 - 4\alpha\gamma}}{2\alpha}. \tag{2.26}$$

With the solution for λ, the solution for the on-diagonal case is complete, resulting in a complete solution for $\theta_{jn'n}$. The solution presented integrates the consensus level vector directly into the estimation process by weighting the importance of each voxel based upon the estimated level of confusion/consensus. There are many benefits to this approach particularly the fact that it provides a performance level estimate that automatically dramatically decreases the importance of high consensus regions. However, because of the integration of the consensus level vector, the estimate for $\theta_{jn'n}$ no longer has a strict statistical interpretation as a measure of sensitivity/specificity.

At this point it is important to clarify the implications of the number of consensus levels. The consensus level estimation for each voxel is integrated into both the estimation of the true segmentation (E-Step) and the performance level parameters (M-Step). For the E-Step, the *probability* that a given voxel belongs to a given consensus level is estimated. Thus, even in the binary consensus level case, the actual state (i.e. *consensus* or *confusion*) exists on a spectrum. To illustrate this, consider a voxel where all raters agree, the probability that this voxel belongs to consensus level $\varsigma = F$ would be estimated to be 1.0 and 0.0 for all other consensus levels. On the other hand, for a voxel where most raters agree on the label but there is some disparity, then, in all likelihood, the probability associated with each consensus level would be greater than zero. For the M-Step, the number of consensus levels indicates the number of possible weighting factors that are taken into account when estimating θ. These weighting factors are then applied to the probability that a given voxel has label *s* and consensus level ς from the E-Step. The optimal number of consensus levels for a given task largely depends upon the difficulty of the labeling task. For a straightforward task where the only confusion about the true label would exist along the

boundary between labels, then the binary consensus level case would be appropriate. For a more difficult problem, such as estimating full brain structure in a multi-atlas multi-label task, more than two consensus levels may be more appropriate and would make for an interesting area of future consideration.

2.5. Initialization Strategy, Convergence Detection, and Model Parameters

The theory presented above provides the framework for the implementation of the COLLATE algorithm. In order to fully implement the algorithm, however, an initialization strategy, method of detecting convergence and the model parameters must be set according to the needs of the application.

1) Initialization: COLLATE can be initialized by either providing an initial estimate of the performance level parameters ($\boldsymbol{\theta}$) or the true segmentation and consensus level vector (usually implemented as an initial estimation of **W**). In this paper, COLLATE is initialized with an initial estimate of **W** as the results of a majority vote algorithm. If the data are available, a probabilistic atlas can be used to provide an initial estimate of the true segmentation [137]. If an initial estimate of the true segmentation is provided, then the iterative process of the E-M algorithm begins by calculating the rater performance parameters from the initial estimate of the true segmentation and consensus level vector.

As opposed to providing an initial estimate of the true segmentation, initial estimates of the rater performance parameters can be provided to initialize COLLATE. Previous algorithms [8] have used this strategy for initialization of the E-M algorithm. If there is no prior information about the performance of the raters then the initialization strategy is generally to assume that all raters are of equally high quality. For example, this could be accomplished by setting the diagonal of $\boldsymbol{\theta}_j = 0.99 \; \forall j$. It should be noted that if an initial estimate of the performance parameters is provided then the iterative COLLATE algorithm would begin with an estimation of the true segmentation.

In all of the simulations and empirical experiments presented in this paper, an initial estimation of the true segmentation is used to initialize the COLLATE algorithm.

2) Convergence: As with all E-M algorithms, the COLLATE algorithm presented in this paper is guaranteed to converge to a local maximum. The detection speedup of convergence is a topic that has

been explored on multiple occasions [138, 139]. The COLLATE algorithm estimates the performance level parameters, the true segmentation and the consensus level given the input data provided by multiple raters. A close monitoring of any of these parameters would provide a quality method of detecting convergence depending upon the application. In this paper, the desired method of convergence detection is through monitoring the change in the performance level parameters. As suggested in STAPLE, the change in the normalized trace of the estimated performance level parameters is the desired method of convergence detection. We use a threshold of $\varepsilon = 1 \times 10^{-3}$ for all simulations and empirical experiments presented in this paper. The normalized trace calculation is given by

$$\frac{1}{LR}\sum_{j=1}^{R} tr(\theta_j). \tag{2.27}$$

The number of iterations required for convergence generally depends upon the number of coverages and the quality of the data passed to the COLLATE algorithm. In this worst case scenario (i.e. low number of coverages, low quality raters) the algorithm generally converges in around 20 iterations in our experience.

3) Data-adaptive Priors: There are three different data-adaptive priors that need to be determined in order to perform the COLLATE algorithm. The first prior that needs to be set is $p(C = 0)$ which describes the probability of a rater giving the incorrect label for a *consensus* voxel. It is of note that in the binary consensus level case, the $p(C = 1) = 1 - p(C = 0)$ and describes the probability that the rater reports the correct label for a *consensus* voxel. In this paper, the values of 0.99 and 0.01 were used for $p(C = 1)$ and $p(C = 0)$, respectively.

The second prior that needs to be set is $f(T_i = s)$. This is equivalent to the prior in the STAPLE algorithm and can be either a global or a spatially varying prior. A spatially varying prior would be optimal in situations when prior knowledge, such as a probabilistic atlas, is available for the given segmentation. In general, this prior can be thought of as simply the probability that voxel *i* has associated label *s*. As a rule of thumb, if explicit spatial information about the true label is available (i.e. information

above and beyond the observed data) it should be integrated into this prior. In all of the simulations presented in this experiment no explicit spatial information is provided, thus, a global prior is used, in which $f(T_i = s)$ is a vector where each element in the vector represents a prior probability for each available label in the segmentation. For notational consistency, we let $\gamma_s = f(T_i = s)$. In a situation where this quantity is not readily available, this value is found using the input segmentations provided by the labelers

$$\gamma_s = \frac{1}{NR}\sum_{j=1}^{R}\sum_{i=1}^{N} I(D_{ij} = s) \tag{2.28}$$

where $I(D_{ij} = s)$ is the indicator function which is equal to 1 when $D_{ij} = s$ and equal to 0 otherwise.

The final prior that needs to be set is $f(C_i = \varsigma \mid T_i = s)$ which indicates the probability that a given voxel is *consensus* or *confusion* given that the true label is *s*. As with the parameter $f(T_i = s)$, this parameter could be a spatially varying prior or a global prior. Again, in all of the simulations and experiments in this paper a global prior is used. In the case of a global prior, let Ψ_i be a binary variable indicating whether or not voxel *i* has been estimated to be *confusion* ($\Psi_i = 0$) or *consensus* ($\Psi_i = 1$). In order to calculate the value of Ψ_i a threshold value of $\tau = 0.95$ is used, which indicates the fraction of raters that need to agree in order for a pixel to be estimated to be consensus. The value of $\Psi_i$is calculated at each voxel by

$$\Psi_i = I\left(\max_s\left(\frac{1}{R}\sum_{j=1}^{R} I\big(D_{ij} = s\big)\right) > \tau\right) \tag{2.29}$$

where s is a value in the set of $\{0, 1, \ldots, L-1\}$.

In addition to the calculation of Ψ_i, the binary variable that estimates the status of each voxel ("consensus" or "confusion"), an estimation of the true label is garnered through the majority vote algorithm at each voxel. Let K_i represent the label estimated through a majority vote algorithm. The value of $\rho_{\varsigma|s} = f(C_i = \varsigma \mid T_i = s)$ is computed by

$$\rho_{\varsigma|s} = \frac{1}{N_s}\sum_{i=1}^{N} I(\Psi_i = \varsigma) I(\mathrm{K}_i = \mathrm{s}) \tag{2.30}$$

where

$$N_s = \sum_{i=1}^{N} I(\mathrm{K_i} = \mathrm{s}). \tag{2.31}$$

It is important to note that there is no guarantee that these methods of calculating the data-adaptive priors are optimal. Nevertheless, these parameters are meant to provide a basis by which COLLATE uses to compute the estimates of the rater performance parameters and the hidden data. Instead of constructing two separate priors (as presented above), it would accomplish a similar goal to use a single spatially varying prior as previous work has suggested [8]. In this case, the estimation of *consensus/confusion* would be integrated implicitly into the spatially varying prior. We felt that exposing two separate priors made the estimation of confusion level and the various opportunities for implementation more explicit. The implementation of the data adaptive priors was meant to be as simple as possible.

3. Methods and Results

3.1. Terminology

In the following results and methodologies presented in this section, several simulations and experiments are presented. These simulations range from a model that matches the COLLATE rater behavior model to an empirical experiment that uses data acquired from human raters. In order for the presentation to be as clear and consistent as possible it is necessary to define some terminology:

- A *label* is an integer valued category assigned to an anatomical location.
- A *rater* is an entity (real or simulated) that reports or observes labels.
- An *observation* is the result of a single rater observing all labels in a given slice (i.e. assigning an integer value to all pixels/voxels in an image).

- A *coverage* is the result of a single rater making exactly one observation of each available slice in the set.
- A *truth model* defines the true labels for all voxels/pixels in all of the available slices. If the slices are the output of a simulation then the truth model is generally known. For empirical data the truth model is the result of an anatomical expert carefully providing a label for each voxel.
- A *generative model of rater behavior* defines the way in which a label fusion algorithm (e.g. STAPLE, COLLATE) models the decision-making process of a given rater.
- The *consensus level vector* is the aspect of the true segmentation that is introduced by the COLLATE algorithm. This vector is fully integrated into the estimation process. The probability that a given voxel is in each consensus level is estimated and this estimated probability is used to determine the weighting of each voxel in the determination of the performance level parameters.
- A *consensus map* is a property specific to the COLLATE algorithm that defines the regions in a specific slice where consensus/confusion is present. This concept is a byproduct of that fact that we are using binary consensus levels, as it dramatically simplifies the representation of the estimated consensus levels. The consensus map is derived from the estimated consensus level vector in the weight variable, $W_{siç}^{(k-1)}$. Mathematically, this is defined as $\sum_{s=0}^{L-1} W_{si0}^{(k)}, \forall i$, where it defines the amount of weight in the in $C = 0$ consensus level. All of the values of the consensus map are $\in [0,1]$, where 0 (black) is referred to as "full consensus" and 1 (white) is referred to as "full confusion." The consensus map can be thought of as the continuous *probability* that each voxel is a voxel of *confusion*.
- A *confusion region* is the collection of high-valued pixels/voxels in the consensus map.
- A *confusion matrix* is an LxL matrix where each element in the matrix defines the probability that a rater would label a voxel with label *s* given that the true label is *s'* in high valued regions of the consensus map. As previously defined, the confusion matrix for rater j is θ_j.

This terminology will be used consistently in the following simulations and empirical experiments.

3.2. Implementation and Evaluation

COLLATE and all simulations were implemented in MATLAB (Mathworks, Natick, MA). The implementation used to produce the results seen in this paper is available via the "MASI Label Fusion" project on the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC, www.nitrc.org/projects/masi-fusion). For all simulations and experiments, both STAPLE and COLLATE use a global prior for the label probabilities (see Eq. (28)). COLLATE uses a global prior for the fraction of each label that are in consensus (see Eq. (30)). Additionally, for all simulations both COLLATE and STAPLE analyze the entire truth model when performing the estimation procedure (i.e. no regions of interest are considered [9]). The empirical data presented in this paper was gathered using the WebMill interface (https://brassie.ece.jhu.edu/Home). For data contributed by human raters, detailed instructions about the labeling procedure were provided to the raters. All human raters were required to perform at least one practice labeling before proceeding to the actual labeling data. All studies were run on a 64 bit quad-core 3.07GHz desktop computer with 13GB of RAM, running Ubuntu 9.04.

When results are presented, the resulting truth model, performance level (confusion matrices) and consensus map estimations are presented. In situations where the true confusion matrix for the model is known, the accuracy of the estimations is presented. Otherwise, a comparison between the resulting STAPLE confusion matrix and COLLATE confusion matrix are presented for visual comparison. In all of the simulations the accuracy of the estimated truth labels is presented for varying numbers of coverages as the fraction of pixels correct in the confusion region. For the empirical simulation the "ground truth" provided by an expert anatomist is provided for visual comparison with the COLLATE truth estimation.

3.3. Simulation 1: Simulation using COLLATE Model of Rater Behavior

The first simulation (Figure II.3) used simulated raters that are nearly identical to the COLLATE model of rater behavior. This means that the consensus map is clearly defined to be low valued outside of the square in the middle of the truth model and high valued inside of the square. The truth model consists of 50 slices of size 100x100 pixels. A collection of 20 simulated raters were created that are described by

confusion matrices with constant valued diagonals. The diagonal values for the raters were linearly spaced between 0.45 and 0.65. Thus, the raters from this simulation were slightly better than chance as there were 5 labels on the truth model. For all slices the size of the confusion region was held constant at 10%.

The purpose of this simulation was to assess the accuracy of COLLATE in a model where there exist well-defined regions of the image where raters are very accurate and other regions where they are only slightly more accurate than chance. The accuracy of truth estimation and the confusion matrices is assessed and compared to the accuracy of STAPLE as a reference point. The accuracy of these estimations was assessed by varying the number of coverages (from 3 to 20) passed to COLLATE. An estimate of the consensus map is provided for eight coverages with 25 Monte Carlo iterations.

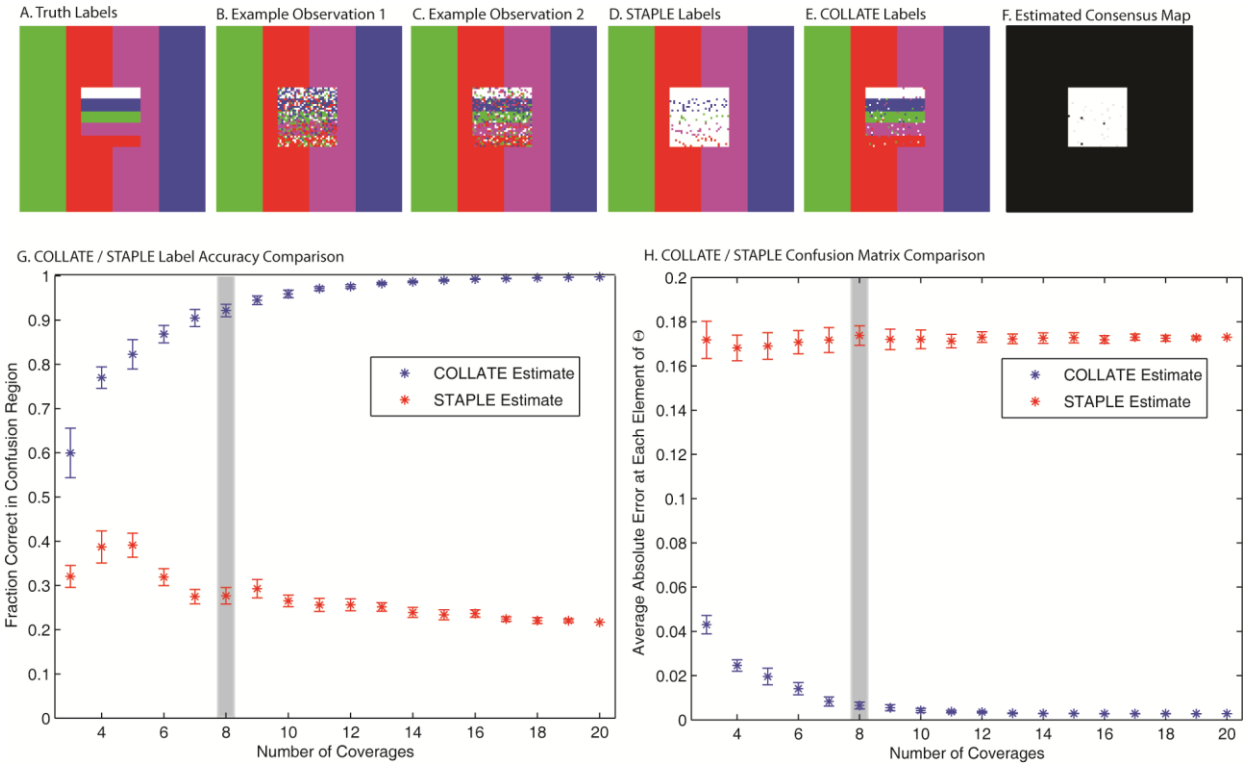


Figure II.3. Results for simulation 1 using the COLLATE model of rater behavior. A representative slice from the truth model can be seen in (A). (B) and (C) represent example observations of the slice seen in (A). The STAPLE estimate using 8 coverages can be seen in (D). The COLLATE estimate using the same observations can be seen in (E). The estimated consensus map can be seen in (F). The accuracy of the estimated labels and confusion matrices can be seen in (G) and (H), respectively. The gray bars seen on (G) and (H) correspond to the number of coverages used in the estimations seen in (D), (E) and (F).

Figure II.3A represents an example slice from the truth model used in this simulation. Note that the 'white' label present in the confusion region (i.e., the light-gray to white area of the consensus map) is not present in the consensus region (i.e. the dark-gray to black area of the consensus map). This was included because of the apparent problem with the STAPLE algorithm for small labels. The reason for this problem is that the limited data used to estimate the performance level parameters for small labels tends to increase the likelihood for label inversion. It is of note that both the parametric prior approach proposed by Commowick et al. [55] and the non-parametric prior proposed by Landman et al. [11] have decreased the likelihood of witnessing label inversion on small labels. Figures II.3B and 3C represent example observation made by the simulated raters with diagonal values of 0.45 and 0.65, respectively. Figure II.3D represents the labels generated by STAPLE using eight coverages. As clearly evident, the presence of the small white label and the poor accuracy of the raters in the confusion region cause STAPLE to converge to an estimate that does not match the truth model. Figure II.3E represents the COLLATE truth estimation which does not suffer from the same failures as the STAPLE estimation. Due to the introduction of the consensus map and the new generative model of rater behavior, COLLATE is able to converge to the correct answer despite the small label and poor labeling accuracy in the confusion region. Figures II.3G and 3H show the accuracy of COLLATE and STAPLE for the simulation with varying numbers of coverages. The results shown in Figure II.3G indicate that the COLLATE truth estimation is consistently more accurate in the confusion region than the STAPLE estimate. Additionally, due to label inversion, STAPLE converges to the incorrect truth estimate while COLLATE converges to the correct truth estimate. The results shown in Figure II.3H show that COLLATE is able to converge to the confusion matrices that match the simulation model, while, not surprisingly, STAPLE converges to a significantly different approximation of the confusion matrices.

3.4. Simulation 2: Data Adaptive Prior Sensitivity

The next simulation (Figures II.4 and 5) was constructed by creating a truth model that is equivalent to the model seen in Figure II.3. The truth model consists of 50 slices of size 100x100 pixels.

A collection of 20 simulated raters were created that were described by confusion matrices with constant valued diagonals. The diagonal values for the raters were linearly spaced between 0.55 and 0.75. Note that despite the low diagonal confusion matrices the raters are still significantly better than chance as there are five labels present on the truth model. Each rater observed one coverage of the truth model.

The purpose of this simulation was to quantify the sensitivity of COLLATE with respect to the $f(C_i = \varsigma \mid T_i = s)$ data adaptive prior. This simulation is broken up into two parts. The first part (Figure II.4) maintains a constant confusion region (50%) and varies the data adaptive prior. The second part (Figure II.5) focuses on the algorithm's ability to estimate the confusion prior by varying the size of the confusion region (5% to 95%). For both parts, the accuracy of the estimated labels is represented as a percent improvement over the STAPLE estimate of the same truth model. A truth model was created such

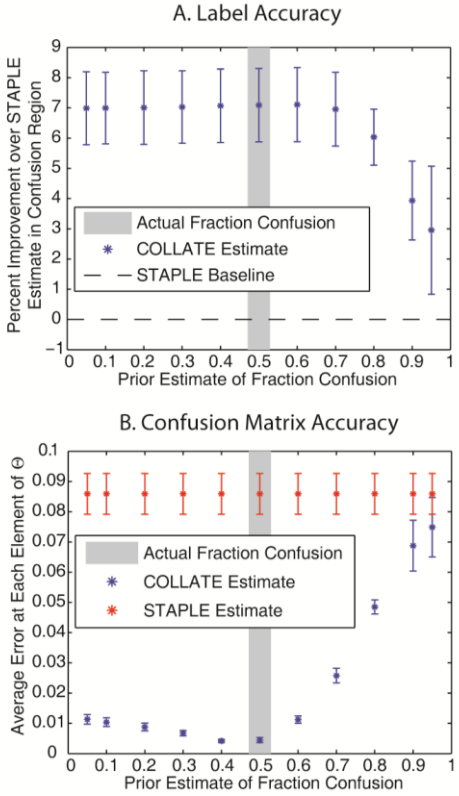


Figure II.4. Results for simulation 2, the COLLATE sensitivity with respect to the estimated confusion region size data-adaptive prior. The sensitivity of the confusion region size prior can be seen in (A) and (B). (A) represents the accuracy of the truth estimation with varying prior estimates from 0.05 to 0.95 for a given confusion region size of 0.5. The accuracy of the truth estimation is presented as a percent improvement over the STAPLE.

that varying the fraction of the image that represents a confusion region is straightforward. For both parts of the simulation a random subset of six raters (six coverages) was chosen to construct the estimates with 10 Monte Carlo iterations.

Figures II.4A and 4B represent the results for a constant 50% confusion region for the truth estimation and confusion matrix accuracy, respectively. These results show that, as expected, the ideal result is obtained when the estimate of the fraction confusion is equal to the actual fraction confusion. An underestimate has little effect on the accuracy of the results for the truth estimation, but causes the confusion matrix accuracy to decrease. An overestimate of the confusion region drastically affects the accuracy of both the truth estimation and the confusion matrix estimation. Nevertheless, regardless of the

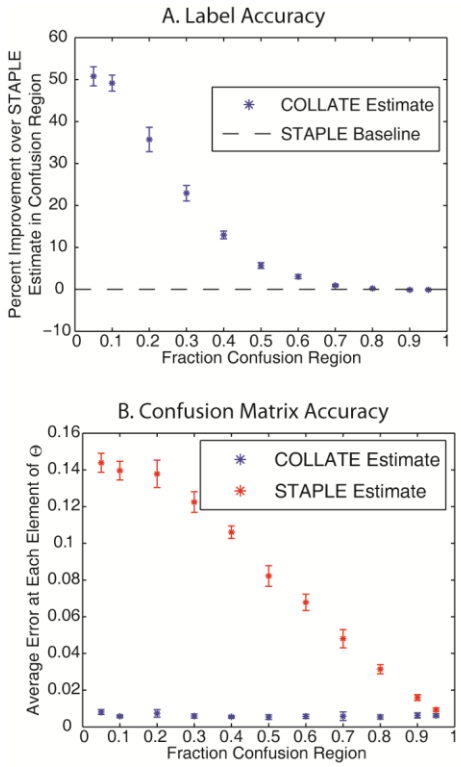


Figure II.5. Results for simulation 2, the accuracy of the COLLATE algorithm with respect to the confusion region size. This tests the ability of the algorithm to estimate the confusion region size. (A) represents the percent improvement for COLLATE over the STAPLE estimation for confusion region sizes varying from 0.05 to 0.95. (B) represents the average absolute error at each element in the confusion matrices for varying confusion region size. Note that the COLLATE estimate accuracy remains constant while the quality of the STAPLE estimate varies depending upon the size of the confusion region. All data presented in this Figure use six coverages for both COLLATE and STAPLE.

data adaptive prior estimate, the COLLATE estimates are consistently better than the estimates obtained by STAPLE. Additionally, confusion matrix accuracy of the STAPLE estimate is consistently different than the confusion matrices used in the simulation as the STAPLE rater behavior model does not take into account the consensus map introduced by the COLLATE algorithm. These results are present on Figure II.3B to emphasize the differences between the two generative models of rater behavior.

Figures II.5A and II.5B represent the results for varying confusion region size. The size is varied between 5% and 95%. The results show that as the confusion region increases toward full confusion, the COLLATE truth estimate and STAPLE truth estimate (Figure II.5A) converge to the same accuracy level. This is due to the fact that if the confusion region represents the entire image, then the COLLATE and STAPLE models of rater behavior are equivalent. Figure II.5B shows that, regardless of the confusion region size, the accuracy of the COLLATE confusion matrix estimates remains approximately constant, while the STAPLE estimate increases in accuracy until a large confusion region is present, in which case the COLLATE and STAPLE confusion matrix estimates are of approximately equal accuracy.

3.5. Simulation 3: Simulation using Boundary Random Raters

The third simulation (Figure II.6) emulates a reliable model of rater behavior by simulating raters that only miss by inaccurately labeling the boundary between two adjacent label regions. This approach is slightly different than previous boundary random rater simulations where each boundary pixel had a 50% chance of being chosen incorrectly. The truth model for this simulation consists of 50 slices of size 100x100 pixels. Once again, a collection of 20 raters were used to observe the truth model slices, however, the raters were designed to only miss at the boundaries and only assign the labels in the adjacent regions to each boundary. This was accomplished by identifying the boundary pixels and applying a "shift" amount (positive or negative) to each boundary pixel. A random number was drawn from a Gaussian distribution, where the standard deviation of the distribution was determined by the quality of the rater. The standard deviations ranged from 1.2 to 3.26 for the best and worst raters respectively.

This simulation assesses the accuracy of the estimates returned by the COLLATE algorithm in a model that closely approximates the way that human raters observe truth models. The accuracy of the truth estimations was assessed for various numbers of coverages, ranging from 3 to 20. Due to the fact that the confusion matrices do not precisely correspond to the proposed generative model of rater behavior, a visual comparison is presented for the COLLATE and STAPLE estimations. Twenty-five Monte Carlo iterations were used. As with the second simulation, an estimated consensus map is provided for eight coverages.

Figure II.6A represents the truth model used for this simulation. Figures II.6B and 6C are representative observations made by a high quality rater and a low quality rater, respectively. Figures

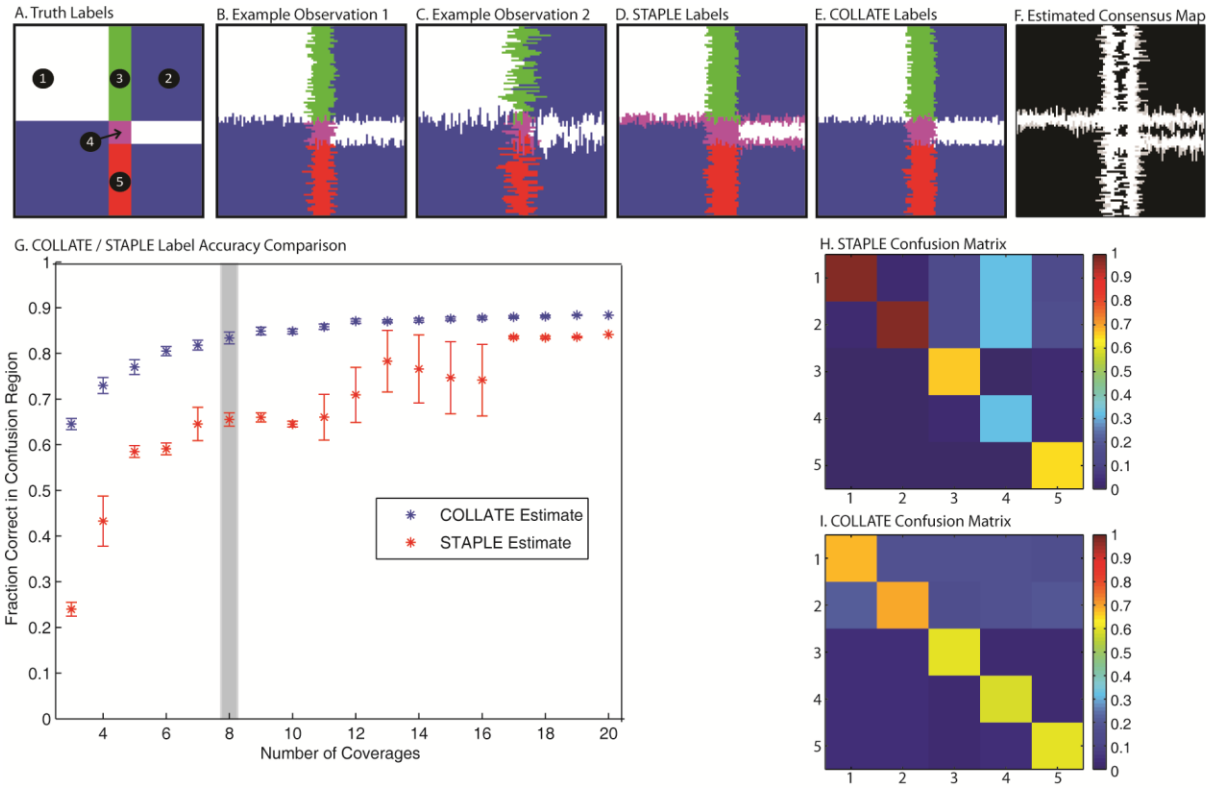


Figure II.6. Results for simulation 3 using boundary random raters. A representative slice from the truth model can be seen in (A). (B) and (C) represent example observations of the slice seen in (A). The STAPLE estimate using eight coverages can be seen in (D). The COLLATE estimate using the same observations can be seen in (E). The estimated consensus map can be seen in (F). The truth estimation accuracy comparison of the two algorithms in the confusion region for varying numbers of coverages can be seen in (G). The gray bar indicates the number of coverages corresponding to the estimates seen in (D), (E). (F), (H) and (I). An example confusion matrix from a single rater from the STAPLE estimate and the COLLATE estimate using eight coverages can be seen in (H) and (I).

II.6D and 6H show the STAPLE estimation of the true labels and an example confusion matrix after 8 coverages. Upon visual inspection it is evident that STAPLE has incorrectly placed the magenta label. The reason for this problem is due to label inversion, which can be seen in the 4^{th} column (which corresponds to the magenta label) of the confusion matrix in Figure II.6H. Label inversion occurs when the confusion matrix estimation indicates when a rater assigns a label other than the intended label. This can have catastrophic effects on the truth estimation as it can cause large regions of the estimation to have incorrect label values. Recent work has focused on the label inversion problem and both parametric and non-parametric priors have been proposed that have been shown to prevent label inversion [11, 55]. On the other hand, the COLLATE estimates of the true labels, consensus map and example confusion matrix can be seen in Figure II.6E, 6F, and 6I, respectively. The COLLATE estimate does not suffer from the same label inversion problem.

COLLATE estimates the rater to have a nearly constant diagonal confusion matrix. This makes logical sense as the raters were not designed to be biased towards certain labels. The reason for this benefit is due to the modified generative model of rater behavior, which serves to normalize the size of the labels in the confusion matrix estimation. In this simulation, the STAPLE confusion matrix estimations largely depend on the size of the region associated with a given label. The reason for this is the fact that larger regions have significantly more consensus voxels. Thus, due to the STAPLE model of rater behavior, the performance level estimations for a given label will be largely dependent on the size of the label in this simulation (see Figure II.6H). On the other hand, COLLATE removes this dependence by weighting the voxels based upon the estimated consensus levels. It is of note that if modifications were made to the traditional STAPLE algorithm (i.e. spatially varying prior, or specifying a region of interest) this dependence may be reduced. The accuracy of the truth estimations with respect to number of coverages is presented in Figure II.6G. As with the second simulation, the COLLATE estimates are of consistently higher accuracy and lower standard deviation than the estimates provided by STAPLE. Note that the y-axis on this plot is the fraction of pixels correct in the confusion region only. The fraction correct would be significantly higher if the consensus regions were included, however, the pixels of

interest in the COLLATE model are the pixels where there is confusion about the true label, and thus, only the pixels in the confusion region are considered.

3.6. Empirical Comparison using Delineations by Human Raters

The empirical experiment presented in Figure II.7 compares the accuracy of the COLLATE and STAPLE algorithms on data generated by human raters. The truth model for the empirical consisted of 10 slices of 70x110 pixels. These slices were selected from a whole-brain scan (182x218x182 voxels) of a healthy individual (after informed written consent) that was cropped to isolate the posterior fossa. A specific region of the brain was isolated (i.e. the posterior fossa) to simplify the labeling process and

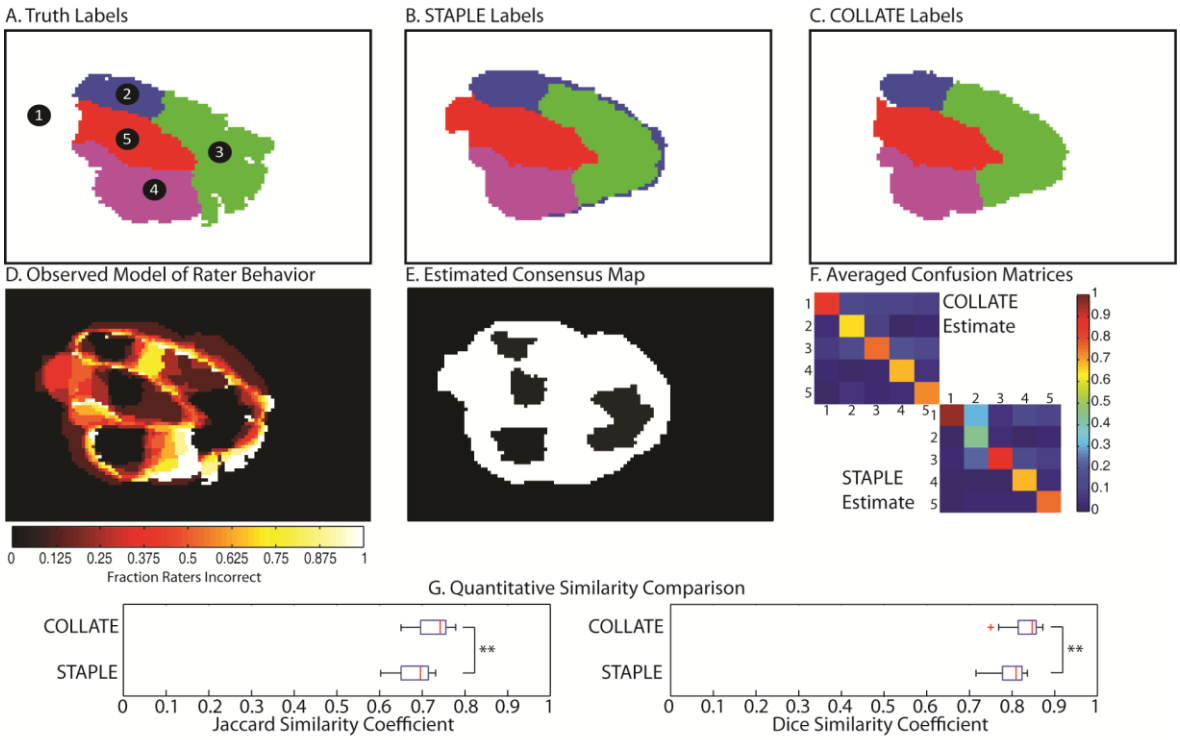


Figure II.7. Empirical experiment using human raters. A representative slice from the 10-slice truth model can be seen in (A). The STAPLE and COLLATE estimates can be seen in (B) and (C), respectively. The observed model of rater behavior can be seen in (D). The color value at each voxel corresponds to the fraction of raters that incorrectly labeled the given voxel. The estimated consensus map can be seen in (E). The averaged confusion matrices for both the STAPLE and COLLATE estimations can be seen in (F). The range of Jaccard Similarity Coefficient values and Dice Similarity Coefficient values can be seen in (G). In both cases, a paired t-test resulted in $p < 0.001$.

allow a large collection of data to be collected in a relatively short amount of time. A collection of eight raters each performed a single coverage using the online WebMill system. The task defined for each rater was to label the sagittal cross-section of a cerebellum. Five different colors were assigned to five different regions of the cerebellum. The color blue was assigned to Lobules I-V (upper lobe), green was assigned to Lobules VI-VII (middle lobe), magenta was assigned to Lobules VIII-X (lower lobe), red was assigned to the Corpus Medulare White Matter and Yellow was assigned to the Vermis. All background pixels were assigned the color white. The raters observed the slices by applying the labels directly to the high resolution Magnetization Prepared Rapid Acquired Gradient Echo (MPRAGE) sequence. While performing each observation a reference image was placed in the top right corner to visually remind the raters of the task to be performed.

The spatial homogeneity and overlap are particularly important when comparing segmentations gathered using clinical data. Thus, the Dice and Jaccard similarity coefficients were used when comparing the accuracy of the truth label estimations acquired from the algorithms. The Dice Similarity Coefficient (DSC) [140]is an often used metric when comparing the spatial overlap between two vectors. The DSC is defined as $(2|A \cap B|)/(|A| + |B|)$ where $|A|$ and $|B|$ represent the area of regions A and B respectively. The Jaccard Similarity Coefficient [141] is another commonly used metric when defining the spatial overlap between two vectors. The Jaccard Similarity Coefficient (sometimes called the Jaccard Index) is defined as $|A \cap B| \; / \; |A \cup B|$ where $|A|$ and $|B|$ have the same meaning as seen in the DSC definition.

Figure II.7 illustrates both that COLLATE can accurately fuse the labels from multiple raters, but also that it can outperform STAPLE. Figure II.7A presents a representative slice from the truth model that was created by an expert neuroanatomist. Figures II.7B and 7C represent the STAPLE and COLLATE estimations of the true labels by fusing the labels from eight raters. Figure II.7B contains several mislabels around the outside boundary of the estimation and inaccurately extends the corpus medulare. It is important to note that these errors are due to partial label inversion. This is a highly studied problem with statistical fusion methods [142, 143] that results from when two labels are commonly confused. This

problem manifests itself in the STAPLE confusion matrix estimate (see the second column of the STAPLE confusion matrix in Figure II.7F).

Figure II.7D represents the observed model of rater behavior and it demonstrates that the raters did not arbitrarily mis-label pixels on the slice. Instead, as previously described, the raters struggled with boundary pixels and regions of low contrast on the original slice. Figure II.7E represents the estimated consensus map, which agrees with the observed model of rater behavior. The estimated consensus map presented here is certainly a little conservative. This could be adjusted by modifying the $p(C = \varsigma)$ prior which is discussed in the theory section. The averaged confusion matrices for both the COLLATE and the STAPLE estimates can be seen in Figure II.7F. Similar to the results seen in the third simulation, the COLLATE averaged estimated confusion matrix is very close to a constant diagonal matrix. This makes intuitive sense, because there is no reason why any given rater would be biased towards a given label value since all labels in this task share a boundary. The averaged STAPLE confusion matrix shows that nearly all raters are perfect at background simply because of the fact that the majority of the pixels are background in the truth model. Additionally, there is near label-inversion in the second column of the averaged STAPLE confusion matrix which explains the presence of the blue label on the perimeter of the green label on the STAPLE truth estimate.

Lastly, the range of Jaccard and Dice Similarity coefficient values for the 10 slices used in this experiment were computed for both the COLLATE estimates and the STAPLE estimates of the true labels. These ranges can be seen in the two plots in Figure II.7G. A paired t-test was performed on both the Jaccard and Dice similarity coefficients and the resulting p-values were found to be less than .001 for both similarity metrics. This indicates that there is significant improvement gained by using COLLATE on empirical data.

3.7. Simulation 4: Simulation using STAPLE Model of Rater Behavior

The fourth, and final, simulation (Figure II.8) presented in this paper assesses the accuracy of the COLLATE algorithm when using a model that matches the STAPLE generative model of rater behavior.

Note that this is equivalent to a COLLATE truth model that has a true consensus map that is all confusion. The truth model consists of 50 slices of 100x100 pixels. The truth model used in this simulation is identical to the model used in Simulation 1 except that the raters miss uniformly throughout the volume. A collection of 20 simulated raters were created that are described by confusion matrices with constant valued diagonals. The diagonal values are linearly spaced from 0.45 to 0.65 for the worst and best rater, respectively.

The purpose of this simulation was to quantify the accuracy of COLLATE when the STAPLE model of rater behavior is fully accurate. As with the previous simulations, the accuracy of the truth estimations and confusion matrix estimations was assessed and compared to the results from the STAPLE algorithm. This was accomplished by varying the number of coverages used to perform the estimations

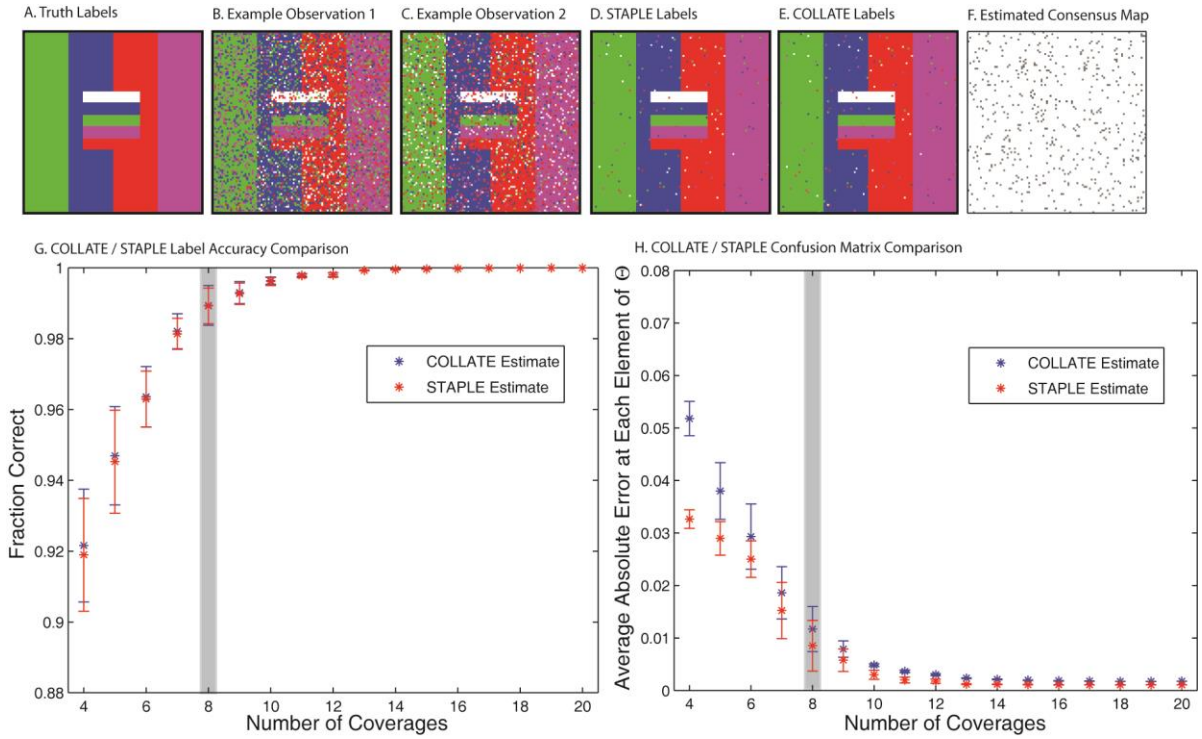


Figure II.8. Results for simulation 4 using STAPLE model of rater behavior. A representative slice from the truth model can be seen in (A) with example observations in (B) and (C). The STAPLE and COLLATE estimates using eight coverages can be seen in (D) and (E), respectively. The estimated consensus map can be seen in (F). The truth estimation accuracy comparison of the two algorithms for varying numbers of coverages can be seen in (G). The confusion matrix accuracy comparison for varying number of coverages can be seen in (H). The gray bars seen on (G) and (H) correspond to the number of coverages used in the estimations seen in (D), (E) and (F).

(from 3 to 20 coverages). Ten Monte Carlo iterations were used to approximate the mean and standard deviation of the estimation accuracy. As with the previous simulations, an estimation of the consensus map is provided for eight coverages.

Figure II.8A shows an example truth model used in the simulation. Figures II.8B and 8C are representative observations of the truth model. Figures II.8D and 8E represent STAPLE and COLLATE truth estimations. Upon visual inspection it appears that they are essentially equivalent. The estimated consensus map can be seen in Figure II.8F. The true consensus map would be fully 'white,' indicating that all of pixels are in the confusion region. By chance, there are several pixels where raters agree, so the consensus map contains several isolated pixels where some level of consensus is present. Figures II.8G and 8H represent the accuracy of the truth estimations and confusion matrix estimations for COLLATE and STAPLE. The average accuracy of the truth estimations (Figure II.8G) by COLLATE is slightly (≈0.005) better than the STAPLE estimations for low numbers of coverages. However, by approximately seven coverages the STAPLE and COLLATE estimates converge to the same level of accuracy. Due to the inaccuracies of the consensus map estimation, the COLLATE estimations of the confusion matrix are less accurate than the STAPLE estimations for all numbers of coverages. Nevertheless, it should be considered that these differences are only on the magnitude of approximately 0.1% for seven or more coverages.

4. Discussion and Conclusion

Herein, we presented an algorithm, COLLATE, for fusing a collection of rater label observations to estimate the consensus level, labeler accuracy and truth labels. COLLATE (1) provides significant improvement over previously developed algorithms, (2) more accurately reflects the realistic rater behavior as seen when human raters segment medical image data, and (3) results in nominal degradation when the rater assumptions are violated (Figure II.7 and 8). Initialization parameters, detection of convergence and other model parameters are clearly defined.

Like its predecessors, COLLATE takes a collection of input observations from a group of raters (human or otherwise) and simultaneously estimates the truth labels and the rater performance parameters ("labeler accuracy"). However, COLLATE also estimates the consensus level of each voxel, which can be viewed as an inherent property of each voxel that determines the likelihood that a given rater would be confused about the label associated with a given voxel. The algorithm presented in this paper is formulated as an instance of the expectation-maximization (E-M) algorithm [116, 138]. As with STAPLE, the decisions of each rater are directly observable, the hidden true segmentation is an integer-valued array corresponding to the label decisions at each voxel. The hidden data in the COLLATE algorithm is augmented with a "consensus level vector" which describes the consensus level of each voxel. The labeler accuracy is iteratively estimated until convergence and is represented in the form of a confusion matrix for each rater. In often used E-M terminology, the *complete* data consists of the rater decisions, which is given, and the true segmentation and "consensus level vector" which are iteratively estimated. In order to perform this estimation, the conditional probability of the *hidden* true segmentation and the "consensus level vector" is evaluated given the rater decisions and the previous estimate of the rater confusion matrices. Convergence to a local maximum is guaranteed. As with previous algorithms, COLLATE is straightforward to apply to medical imaging data acquired from human raters or automated algorithms.

The sensitivity of the consensus-based priors was analyzed for varying confusion region sizes and estimates of the confusion region size. The sensitivity was captured by comparing the accuracy of both the COLLATE truth estimate and confusion matrix parameters to that of STAPLE. Optimal estimations are obtained when priors match the truth model; however, our analysis also indicates that when the confusion region size prior is underestimated the accuracy of the results are significantly better than when the confusion regions size prior is overestimated.

COLLATE is able to accurately estimate the quality of the raters and does not suffer from commonly encountered problems with STAPLE, such as label inversion. For small confusion region sizes, COLLATE significantly outperforms STAPLE for both the truth label accuracy and the confusion matrix accuracy. For large confusion region sizes, COLLATE and STAPLE converge to the same

accuracy level. The benefits of the estimated "consensus level vector" are demonstrated in the differences in the estimated confusion matrices — which reflect a more "physical" characterization of failure likelihood. In the COLLATE estimates, likelihood of error is not biased by the area of the region being labeled.

In this paper, both COLLATE and STAPLE utilize strictly global priors when estimating the true segmentations and the performance level parameters. However, one could use a spatially varying prior instead of a global prior for $f(T_i = s)$. A spatially varying prior has the potential to prevent some of the poor STAPLE segmentation estimations that are presented in this paper (e.g. Figure II.6). Nevertheless, simply implementing a spatially varying prior does not make COLLATE and STAPLE equivalent. COLLATE makes an iterative estimate of the consensus level of each voxel. Voxels that are in high consensus are de-weighted when calculating the performance level parameters. Regardless of whether a global or spatially varying prior was used, STAPLE would consider all voxels to have an equal impact on the calculation of the performance level parameters. Thus, if a given label is present in multiple consensus levels (such as Figure II.3) the final STAPLE estimation of the performance level estimations would be artificially high in the regions where there is significant confusion.

Another modification to STAPLE, proposed by Rohlfing et al.[9], is to select only voxels that are not in consensus when performing the statistical fusion. In some cases, selecting a subset of the total voxels makes the STAPLE model of rater behavior significantly more appropriate and accurate. For the binary consensus level case seen in this paper, the dramatic improvement by COLLATE over STAPLE would certainly be lessened. In some cases, the regions of consensus and confusion are clearly defined and easily detected (e.g. the simulation presented Figure II.3). In this scenario, COLLATE with binary consensus levels is essentially equivalent to performing STAPLE only over the confusion region. Nevertheless, COLLATE provides a framework for integrating any number of consensus levels directly into the estimation process without the need to perform any pre-processing to determine a reasonable region of interest for processing. This framework is the primary contribution of this paper and we feel that this new perspective on the problem will provide fascinating avenues for continuing research.

CHAPTER III

FORMULATING SPATIALLY VARYING PERFORMANCE

1. Overview

Optimality of statistical fusion frameworks hinges upon the validity of the underlying stochastic model of how a rater errs (i.e., labeling process model). Existing methods to simultaneously estimate rater performance (e.g., STAPLE approaches) have used spatially invariant models (i.e. the probability of error does not change voxel-by-voxel) [8, 11, 50, 58, 144]. On the other end of the spectrum, voting based techniques (including global [48, 58, 59] and local [57, 59, 61] approaches) ignore the spatial relation between voxels and fuse each voxel independently. However, these models of observation behavior are at odds with an intuitive notion of rater performance. Regardless of the fusion context (e.g. human raters or multiple atlases), it would not be surprising if the manner in which the labels were observed resulted in spatially varying performance. For example, in a multi-atlas context (Figure III.1), there exist regions where the quality of registration in a multi-atlas based approach is better than others, and, not surprisingly, this quality level is generally smooth on a semi-local level. As a result, we are left with several issues in the field of label fusion, primarily we lack, i) an alternative between the "global" STAPLE method and "local" voting techniques and ii) an understanding of the observed disparity between STAPLE and majority vote on multi-atlas segmentation. By addressing these two issues, a statistical fusion framework would more accurately model the manner in which labels were observed, and, hopefully, result in robustly fused segmentation estimates, regardless of the context.

Herein, we propose an extension to the STAPLE approach, "Spatial STAPLE," to seamlessly account for spatially varying performance. This is accomplished by extending the performance level parameters present in STAPLE to a voxelwise *performance level field* that is unique to each rater. By

estimating a field instead of global parameters, Spatial STAPLE captures the spatially varying behavior that is often present in 1) the fusion of human raters, 2) multiple segmentation algorithms and 3) multiple registered atlases. Additionally, as will become evident later in this manuscript, these performance level fields are guaranteed to be smooth, which allows for a seamless semi-local approach to the fusion of information. This approach is validated for both simulated an empirical datasets modeling both the fusion of human raters and a multi-atlas context. The results suggest that Spatial STAPLE provides a valuable framework that can be utilized to construct an optimal framework for fusion of observed labels, regardless of the context.

It is important to note that, while Spatial STAPLE is demonstrated to perform very well on a variety of label fusion problems, it is, at its heart, a model of human observation behavior. Like STAPLE, Spatial STAPLE 1) models the raters as a group of collectively unbiased observers and 2) does not integrate intensity directly into the estimation process. As a result, a claim that Spatial STAPLE is consistently applicable to a multi-atlas context requires further validation. Herein, we demonstrate that Spatial STAPLE is capable of outperforming multi-atlas fusion methods (e.g., a locally weighted vote) for a CT segmentation application. However, it is unlikely that Spatial STAPLE would be able to outperform multi-atlas fusion techniques for problems in which intensity information provides highly valuable information about the complex relationships between labels and intensity (e.g., whole brain segmentation

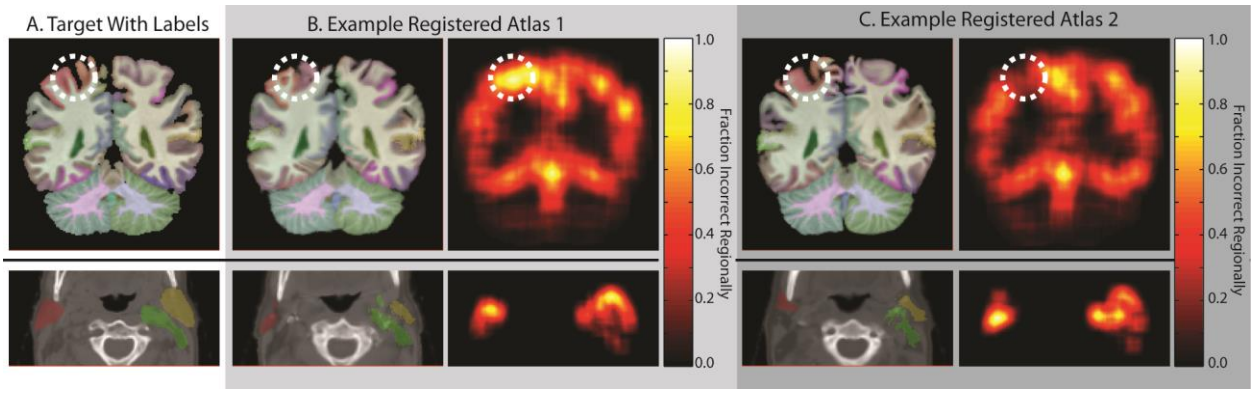


Figure III.1. Registered atlases exhibit spatially varying behavior. Representative slices from an expertly labeled MR brain image and CT head and neck image are shown in (A). Example registered atlases with their local performance can be seen in (B) and (C). Note that atlases exhibit smooth spatially varying performance that is unique to each atlas.

[48, 49, 59]).

Similar techniques to the ones presented in this paper have been proposed. Sabuncu, *et al* [59] proposed accounting for semi-local performance by augmenting locally weighted vote to include a Markov Random Field (MRF). However, this technique is particularly sensitive to using images whose intensities are normalized to one another. This manuscript extends an initial conference publication of the same underlying theory [49] with additional derivations, discussion, experiments and extensions. Subsequently, block-wise neighborhoods were proposed within a meta-algorithm for fusion of local classifiers [67] based upon a Maximum *a posteriori* (MAP) STAPLE framework [55]. Here, we focus specifically on the *performance level field* theory for statistical fusion rather than statistical boosting/meta-analysis.

This chapter is organized in the following manner. Section 2 describes the Spatial STAPLE algorithm and discusses initialization and implementation. Section 3 compares Spatial STAPLE to traditional STAPLE, majority vote and locally weighted vote on a series of experiments and simulations. Lastly, Section 4 provides additional discussion and brief concluding remarks.

2. Theory

The following derivation of Spatial STAPLE closely follows previous derivations of fusion frameworks [8].

2.1. Problem Definition

Consider an image of N voxels with the task of determining the correct label for each voxel in that image. Consider a collection of R raters that provide an observed delineation for each of N voxels exactly once. The index variable *i* will be used to iterate over the N voxels and the index variable *j* will be used to iterate over the R raters. The set of possible labels, **L,** represents the set of possible values that a rater can assign to all N voxels. Let **D** be an N x R matrix describing the labeling decisions of all R raters

at all N voxels where $D_{ij} \in \{0, 1, \ldots, L-1\}$. Let $\mathbf{T}$ be a vector of N elements that represents the hidden true segmentation for all voxels, where $T_i \in \{0, 1, \ldots, L-1\}$.

In the traditional rater model presented in[8], the raters' quality of observation is characterized by $\boldsymbol{\theta}$, the performance level parameters for all raters. In this model, each element, $\boldsymbol{\theta}_j$ (i.e., the performance level parameters for rater j) is an $L \times L$ confusion matrix, where each element in the matrix, $\boldsymbol{\theta}_{js's}$ represents the probability that rater j would observe label s' given that the underlying true label is s. These performance level parameters are global parameters that are utilized at all voxels in order to obtain the estimate of the true segmentation.

Here, we extend these performance level parameters to characterize the performance of each rater with respect to spatial position. As a result, we estimate a *performance level field* for each rater, where each element of $\boldsymbol{\theta}$, $\boldsymbol{\theta}_j(\vec{\boldsymbol{x}})$, represents the performance level (or confusion matrix) associated with rater j at spatial coordinate $\vec{\boldsymbol{x}}$. Additionally, we define $\boldsymbol{B}(\vec{\boldsymbol{x}})$ to be the pooling region (i.e. the spatial region) over which $\boldsymbol{\theta}_j(\vec{\boldsymbol{x}})$ is influenced. We simplify this construct such that the performance level field is discretely defined at every voxel, $\boldsymbol{\theta}_j(\vec{\boldsymbol{x}}) \rightarrow \boldsymbol{\theta}_{ji}$ and the pooling regions is defined as a collection of voxels given by $\boldsymbol{B}(\vec{\boldsymbol{x}}) \rightarrow \boldsymbol{B}_i$. Figure III.2 illustrates the performance level field representation.

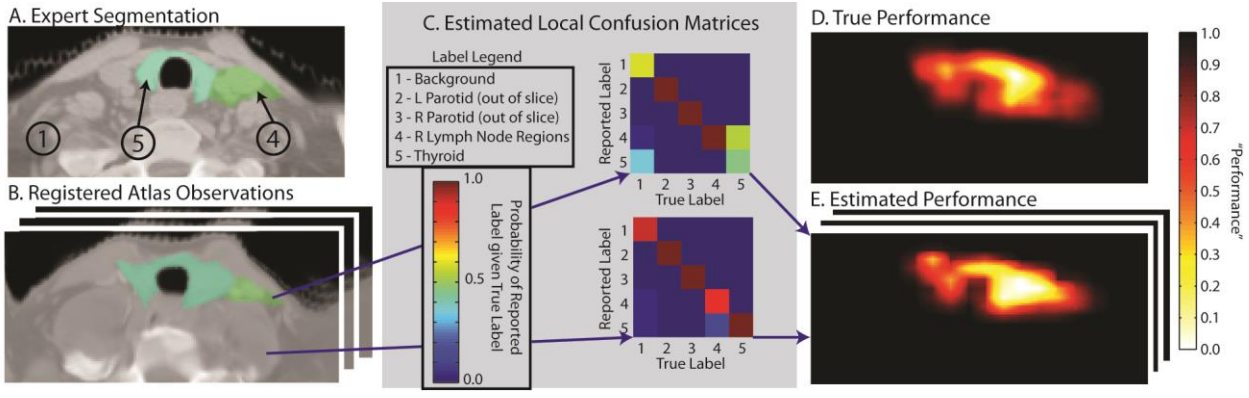


Figure III.2. Demonstration of the Spatial STAPLE performance level field estimation procedure. An example expert segmentation can be seen in (A) with a collection of registered atlas observations seen in (B). Spatial STAPLE estimates local confusion matrices (C) in order to construct a whole-image estimate of performance that is smooth and spatially varying. The true performance for the atlas seen in (B) can be seen in (D) and the estimated performance from Spatial STAPLE presented in (E). Note that the intensity in (E) is an indication of average "performance" – i.e., the average diagonal element of Θ.

2.2. Spatial STAPLE Algorithm

The goal of Spatial STAPLE is to accurately estimate the performance level field of the R raters given the rater segmentation decisions and the estimation of the truth. The estimated performance level field will be calculated such that it maximizes the complete data log likelihood function

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ln f(\boldsymbol{D}, \boldsymbol{T}|\boldsymbol{\theta}) \tag{3.1}$$

It is assumed that the segmentation decisions are conditionally independent given the true segmentation and the performance level parameters, that is $(D_{ij}|T_i\theta_j) \perp (D_{ij'}|T_i\theta_{j'})\forall \mathrm{j} \neq \mathrm{j}'$. This model expresses the assumption that the raters derive their segmentations of the same truth model independently from one another and that the quality of the segmentations is captured by the estimation of the performance level field. The estimated performance level parameters for a given rater at differing voxels are not necessarily conditionally independent.

Our version of the expectation-maximization (E-M) algorithm used to solve (1) is now presented. The complete data used to solve this E-M algorithm is the observed data, **D**, and the true segmentation of each voxel **T**. The true segmentation **T** is regarded as the missing or hidden data, and is unobservable. Let $\boldsymbol{\theta}_{ji}$ be the covariance, or confusion, matrix associated with rater j at voxel i and let

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_{11} & \boldsymbol{\theta}_{21} & \dots & \boldsymbol{\theta}_{R1} \\ \boldsymbol{\theta}_{12} & \boldsymbol{\theta}_{22} & \dots & \boldsymbol{\theta}_{R2} \\ \dots & \dots & \dots & \dots \\ \boldsymbol{\theta}_{1N} & \boldsymbol{\theta}_{2N} & \dots & \boldsymbol{\theta}_{RN} \end{bmatrix} \tag{3.2}$$

be the complete set of unknown parameters for the R segmentations. Let $f(\boldsymbol{D}, \boldsymbol{T}|\boldsymbol{\theta})$ denote the probability mass function of the random vector corresponding to the complete data. The complete data log likelihood function is presented as $\ln L_c\{\boldsymbol{\theta}\} = \ln f(\boldsymbol{D}, \boldsymbol{T}|\boldsymbol{\theta})$. We apply E-M to iteratively estimate and maximize the complete data log likelihood function. Let k denote the iteration for which all estimates were obtained. For more detail on E-M in the statistical fusion model see [8, 50].

2.3. E-Step: Estimation of the Voxelwise Label Probabilities

We first derive an expression for the conditional probability density function of the true segmentation at each voxel given the raters decisions and the previous estimate of the performance fields, $W_{si}^{(k)} \equiv f(T_i = s|\boldsymbol{D_i}, \boldsymbol{\theta}^{(k-1)})$. Applying Bayes' rule and the fact that all of the observations are conditionally independent we obtain a MAP formulation of the underlying segmentation

$$W_{si}^{(k)} = \frac{f(T_i = s)\prod_j f\left(D_{ij} = s' \middle| T_i = s, \boldsymbol{\theta}_{ji}^{(k-1)}\right)}{\sum_n f(T_i = n)\prod_j f\left(D_{ij} = s' \middle| T_i = n, \boldsymbol{\theta}_{ji}^{(k-1)}\right)} \tag{3.3}$$

where $f(T_i = s)$ represents an *a priori* estimate of the true segmentation and $\boldsymbol{\theta}_j^{(k-1)}$ is the prior estimate of the performance level fields. The distribution $f(T_i = s)$ will be discussed more thoroughly later in this manuscript. Finally, as will be seen in the calculation of the performance level fields, the distribution given by $f\left(D_{ij} = s' \middle| T_i = s, \boldsymbol{\theta}_j^{(k-1)}\right)$ simplifies directly to $\boldsymbol{\theta}_{jis's}^{(k-1)}$. Thus, the final equation for $\boldsymbol{W}$ is given by

$$W_{si}^{(k)} = \frac{f(T_i = s)\prod_j \boldsymbol{\theta}_{jis's}^{(k-1)}}{\sum_n f(T_i = n)\prod_j \boldsymbol{\theta}_{jis'n}^{(k-1)}} \tag{3.4}$$

This estimation of the conditional expectation of the complete data log likelihood function is almost identical to the STAPLE approach. The major difference is the utilization of the *performance level field* (i.e., $\boldsymbol{\theta}_{jis's}^{(k-1)}$) as opposed to the global performance level parameters as seen in [8].

2.4. M-Step: Estimation of the Performance Fields via Maximization

Given the estimated weight variable, $W_{si}^{(k-1)}$, which represents the conditional probability that the true segmentation of voxel i is equal to label s, it is now possible to estimate the rater performance field that maximizes the conditional expectation of the complete data log likelihood function. Considering each rater and voxel separately, we find the field estimates $\boldsymbol{\theta}_{ji}^{(k)}$ by iterating only over the voxels given by pooling region $\boldsymbol{B}_i$

$$\boldsymbol{\theta}_{ji}^{(k)} = \arg\max_{\boldsymbol{\theta}_{ji}} \sum_{i' \in \boldsymbol{B}_i} E\left[\ln f\left(D_{i'j} \middle| T_{i'}, \theta_j\right) | \boldsymbol{D}, \boldsymbol{\theta}_j^{(k-1)}\right] \tag{3.5}$$

where, $\boldsymbol{\theta}_{ji}^{(k)}$ is only defined over voxels $i' \in \boldsymbol{B}_i$. Carrying out the expectation yields

$$\boldsymbol{\theta}_{ji}^{(k)} = \arg\max_{\boldsymbol{\theta}_{ji}} \sum_{s'} \sum_{i' \in \boldsymbol{B}_i : D_{i'j} = s'} \sum_{s} W_{si'}^{(k)} \ln \theta_{jis's}^{(k-1)} \tag{3.6}$$

At this point, we are left with the task of maximizing all of the elements in the performance level field using the constraint $\sum_{s'} \theta_{jis's} = 1$ which can be solved with a Lagrange multiplier approach. The final solution for $\theta_{jis's}^{(k)}$ is given by

$$\theta_{jis's}^{(k)} = \frac{\sum_{i' \in \boldsymbol{B}_i : D_{i'j} = s'} W_{si'}^{(k)}}{\sum_{i' \in \boldsymbol{B}_i} W_{si'}^{(k)}}. \tag{3.7}$$

2.5. Accounting for Limited Data and Computational Concerns

Particularly in multi-atlas based segmentation, the spatial quality of an observed segmentation can vary dramatically in a relatively small region (Figure III.1). As a result, the number of voxels in a given pooling region, $\boldsymbol{B}_i$, should be relatively small in order to accurately characterize these semi-local performance variations. However, if the number of voxels contained in a given pooling region is too small, the performance level field will be unstable due to limited data. Additionally, if a given label is not observed (or only observed a handful of times) in the spatial region $\boldsymbol{B}_i$, then it will be impossible to estimate the related elements in the associated confusion matrix $\theta_{jis's}$. Thus, we introduce the idea of using a whole-image estimate of the performance level parameters for regularization.

For computational and stability concerns, we introduce an implicit prior in the following form. Let $\boldsymbol{\theta}_j^{(0)}$ be the confusion matrix associated with rater *j* estimated from an appropriate algorithm. The estimation of the performance level parameters seen in (13) would then be reformulated to be

$$\theta_{jis's}^{(k)} = \frac{\sigma_{ijs}\theta_{js's}^{(0)} + \sum_{i' \in \boldsymbol{B}_i : D_{i'j} = s'} W_{si'}^{(k)}}{\sigma_{ijs}\sum_{s} \theta_{js's}^{(0)} + \sum_{i' \in \boldsymbol{B}_i} W_{si'}^{(k)}} \tag{3.8}$$

where $\sigma_{ijs'}$ is a scale factor that is dependent upon the size of the pooling region, $\boldsymbol{B}_i$, rater j and label s'. Our empirically derived expression for this scale factor is

$$\sigma_{ijs'} = I\left(N_{ijs'} < \frac{|\boldsymbol{B}_i|}{L}\right)\left(\frac{|\boldsymbol{B}_i|}{L} - N_{ijs'}\right)\kappa \tag{3.9}$$

where $I(\cdot)$ is the indicator function, $N_{ijs'}$ is the number of times rater j observed label s' in pooling region $\boldsymbol{B}_i$, $|\boldsymbol{B}_i|$ represents the cardinality of the pooling region (i.e. the number of voxels in the region) and κ is a scalar constant. Unless otherwise noted, the value of κ is unity for all presented experiments. This factor adjusts the impact of the implicit global estimate of performance on the estimate of performance for given rater j, label s' and voxel i.

Numerous approaches could be used to construct the performance level prior (e.g. STAPLE [8], COLLATE [50], Majority Vote, Locally Weighted Vote, etc..). In general, for the fusion of human raters (i.e. when no intensity information is available), we use STAPLE. For multi-atlas segmentation we use Majority Vote while ignoring consensus voxels [144].

The approaches presented in (8) and (9) utilize the observed segmentations to construct a non-parametric estimate of the underlying global performance level parameters. Alternatively, a parametric approach could be used to provide more stability in the performance level field, e.g., [55] used a component-wise beta distribution for the performance level parameters. To date, parametric methods have neglected the interdependence of the distribution of individual entries in the performance level matrix; hence, we advocate non-parametric approaches. Formulating the estimation of performance level parameters optimally in a maximum *a posteriori* framework remains an open problem.

Moreover, calculating confusion matrixes for all raters at all voxels is daunting challenge from both a computational and resource perspective. Hence, we seek to sample and interpolate this field to reduce algorithm complexity. The sample locations are referred to as *seed points*. Herein, we apply linear interpolation using a rectilinear grid.

In this context, implementation of Spatial STAPLE's performance level field presented in this paper can be interpreted as a collection of sliding windows with varying levels of overlap. Herein, we use

the same size window for all seed points and report this size as a fraction of the field of view in each cardinal direction. In other words, a window size of 0.1, would represent a window that is $0.1X \times 0.1Y \times 0.1Z$, where X, Y, and Z are the length of each of the dimensions on the input image. Lastly, the amount of overlap between windows is reported as the fraction linear overlap along each of the principle directions, i.e., an overlap of 0.5 would indicate that there is 50% overlap between consecutive windows along the X, Y, and Z directions.

2.6. Initialization Strategy, Convergence Detection, and the Prior Distributions

1) Initialization: Spatial STAPLE may be initialized by either providing an initial estimate of $\boldsymbol{\theta}$ or $\boldsymbol{W}$. Herein, Spatial STAPLE is initialized with an initial estimate of $\boldsymbol{W}$ as the results of a majority vote algorithm. Alternatively, an initial estimate of $\boldsymbol{\theta}$ could be provided, however, if initialized in the manner suggested in [8] then it is essentially the same as initializing $\boldsymbol{W}$ to a majority vote estimate.

2) Convergence: Spatial STAPLE is guaranteed to converge given its use of E-M [138, 139]. In this implementation, we detect convergence by monitoring the change in the trace of the confusion matrix estimates at each of the seed points. We use a threshold of $\varepsilon = 1 \times 10^{-5}$ for all simulations and empirical experiments presented in this paper. In the worst case empirical trials presented here (i.e. low number of raters, low quality raters) the algorithm converged in less than 20 iterations.

3) Prior Distribution: As with STAPLE, the *a priori* distribution, $f(T_i = s)$, must be defined. This can range from a global parameter to a spatially varying prior. As with the STAPLE implementation, we let $\gamma_s = f(T_i = s)$ and define it as the empirically observed label frequencies for a most general definition. When available, intensity information can be integrated into this prior. In our case, this spatially varying prior would be defined in a manner that is identical to a log-odds majority vote [59]. We let $\psi_{is} = f(T_i = s)$ and define:

$$\psi_{is} = \frac{1}{Z}\sum_{j=1}^{R} e^{\rho \tilde{D}_{ji}^{l}} \tag{3.10}$$

where Z is the partition function, $\widetilde{D}_{ji}^{l}$ is the value of the signed distance transform of the observations of rater j and voxel i for label l. The parameter $p > 0$ defines the slope constant of the log-odds type approach (herein, we use $p = 1$). Finally, for simplicity we define the final value of $f(T_i = s)$ in terms of both the global prior and the spatially varying prior

$$f(T_i = s) = \alpha\, \psi_{is} + (1 - \alpha)\, \gamma_s \tag{3.11}$$

where $\alpha \in [0,1]$ is a parameter that creates a continuum between local and global approaches to prior specification. In general, for STAPLE and Spatial STAPLE we use a prior governed by $\alpha = 1$.

3. Methods and Results

3.1. Implementation and Evaluation

We present four experiments: i) a simulation modeling human behavior, ii) an empirical fusion of human raters for the segmentation of malignant glioma, iii) a simulation modeling the fusion of multiple whole-brain segmentation algorithms using multiple results from a locally weighted vote, and iv) an empirical, multi-atlas based experiment using expert-labeled head and neck CT scans. All experiments were implemented in MATLAB (Mathworks, Natick, MA); complete source code is available via the "MASI Label Fusion" project on the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC).

3.2. Simulation using a Boundary Model of Human Behavior

First, we compared the statistical fusion methods using a model of human behavior by simulating raters that only miss by inaccurately labeling boundary voxels (Figure III.3A and B, similar to [50]). The truth model consisted of a 3-D volume ($80 \times 60 \times 60$) with 15 embedded labeled cubes. The truth model was observed by 16 different raters, where each rater was "perfect" in one sixteenth of the total volume and exhibited boundary error behavior in the remaining regions. In the regions exhibiting boundary errors, the level of boundary shift for each boundary voxel was chosen from a Gaussian distribution with zero

mean and standard deviation — distributed $U(2.5,3.5)$. When all 16 raters are considered, there exists a single rater that is perfect in each of the sub-regions.

1) Overall Accuracy Comparison: Accuracies of Spatial STAPLE, STAPLE and majority vote were assessed with 10 Monte Carlo iterations for each of a varying numbers of raters (ranging from 5 to 16). Spatial STAPLE was applied with: a 0.15 window size fraction (i.e. $12 \times 9 \times 9$), an overlap ratio of 0.5, linear interpolation, and a global performance estimate based on STAPLE with a value of $\kappa = 1$ (see Eq. 9).

The results from this simulation can be seen in Figure III.3C-F. The performance of Spatial STAPLE along the boundaries between the various regions is qualitatively superior to the other methods, as seen by comparing Figures III.3C-3E. Quantitative results demonstrate that increasing the numbers of raters increases the benefit of Spatial STAPLE over both STAPLE and majority vote (Figure III.3F).

2) Sensitivity to the Model Parameters: In addition to an overall accuracy comparison, we assess the sensitivity of Spatial STAPLE to the i) the global performance level bias amount (κ) ii) the size of

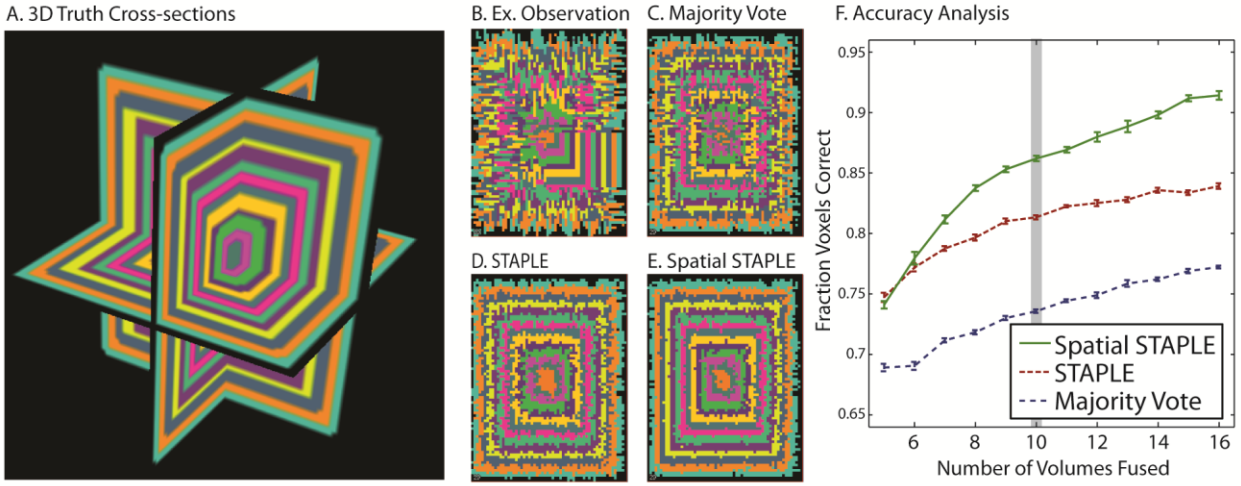


Figure III.3. Results for the human rater simulation. The cross-sectional view of the truth model used in this simulation can be seen in (A). An example observation utilizing the boundary model of human behavior can be seen in (B). Note the fact that there exists a unique region where each rater is perfect in these observations. The corresponding label estimate from Majority Vote, STAPLE and Spatial STAPLE can be seen in (C)-(E), respectively. All displayed estimates were constructed using 10 raters. Lastly, an accuracy analysis can be seen in (F), note that with increasing volumes, Spatial STAPLE continually outperforms both STAPLE and Majority Vote.

each window in the calculation of the performance level field calculations and iii) the amount of overlap between the various windows in the field calculations (Figure III.4). Other than the swept parameter of interest, Spatial STAPLE was applied with a 0.15 window size fraction, an overlap ratio of 0.5, and a value of $\kappa = 1$. The results from these experiments are reported as the percent improvement over STAPLE (in terms of fraction voxels correct).

The global performance level bias (κ, see Eq. 9) was swept in 21 logarithmic steps between 10^{-3} to 10^{3} (Figure III.4A). For both very low and very high bias amounts the accuracy of Spatial STAPLE degrades. However, as the amount of bias approaches unity we see that Spatial STAPLE is vastly superior to STAPLE. This level of improvement plateaus for approximately two orders of magnitude indicating that the accuracy of the algorithm is not particularly sensitive to the strength of the biasing prior. Note that as the strength of the bias increases, the accuracy of Spatial STAPLE converges to the accuracy level of STAPLE due to the fact that the prior estimate of the performance level parameters are estimated from the STAPLE algorithm.

The window size parameter was swept in 20 linear steps between 0.05 (~1-2 voxel windows) and 0.95 (essentially global fusion) (Figure III.4B). For very small window sizes (i.e. 0.05) the accuracy of Spatial STAPLE is worse than STAPLE due to the inability to estimate accurate performance using such

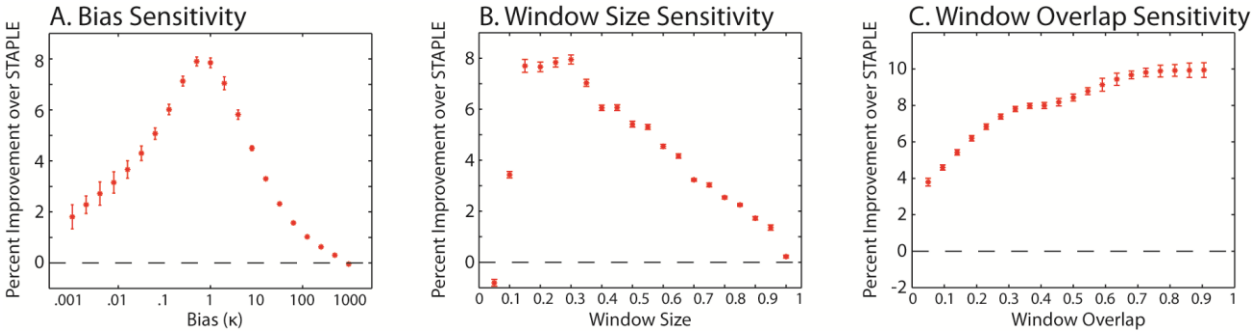


Figure III.4. Assessment of Spatial STAPLE sensitivity with respect to various model parameters for the human rater simulation. For each plot the percent improvement exhibited by Spatial STAPLE over STAPLE is assessed. The plot seen in (A) indicates the sensitivity of Spatial STAPLE to the impact of the global estimate of the performance level parameters. (B) indicates the sensitivity to the size of the pooling region (or window) associated with the voxelwise performance estimate. Lastly, plot (C) indicates the sensitivity to the amount of overlap between windows. The window overlap is a proxy for the number of seed points used in the estimation of the performance level field.

few voxels. However, for relatively small window sizes (i.e. between 0.15 and 0.3), the accuracy of Spatial STAPLE is at a maximum. For window sizes beyond 0.2, the accuracy slowly decreases and converges to an accuracy level that is approximately equal to that of STAPLE.

The window overlap parameter was swept in 20 linear steps between 0.05 (very little overlap) and 0.925 (~3-4 voxel difference) (Figure III.4C). For increasing amounts of overlap the accuracy of Spatial STAPLE increases, as the regional performance level parameters more accurately reflect the area surrounding the voxel of interest. However, beyond 0.5 overlap the increase in accuracy is relatively small. Additionally, as the amount of overlap increases the computational time significantly increases as well.

3.3. Empirical Fusion for Segmentation of Malignant Gliomas

Second, we analyze the accuracy of majority vote, STAPLE, and Spatial STAPLE on the fusion of multiple human raters for the segmentation of malignant gliomas. The cancer patients utilized in this paper are a collection of 15 pre-operative T1-weighted brain MRI scans based on varied (but standard of care) imaging protocols that were obtained in anonymous form under IRB approval. The resolution of

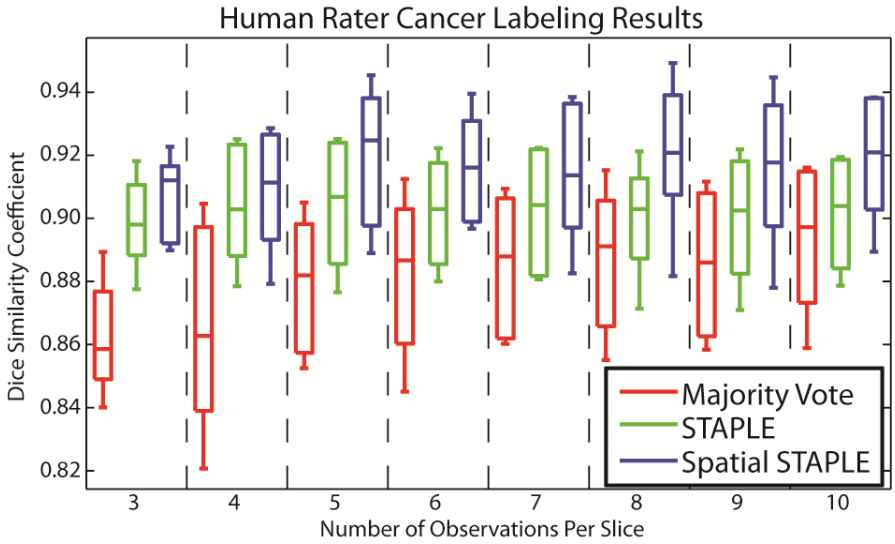


Figure III.5. Qualitative results for the human rater cancer label experiment. The accuracy of majority vote, STAPLE and Spatial STAPLE are considered with varying numbers of observations per slice (or "coverages"). For all number of observations per slice, Spatial STAPLE exhibits statistically significant improvement over both majority vote and STAPLE.

each of the cancerous brains was $1 \times 1 \times 3\ \text{mm}^3$. The corresponding "ground truth" labels associated with each of the cancerous brains were obtained from an expert labeler. A collection of 60 minimally trained undergraduate students provided labels on a slice-by-slice basis (with the students providing anywhere between approximately 10 and 250 observed slices).

The results for the fusion of 8 of these cancerous brains are presented in Figures III.5 and 6. The remaining 7 volumes were used as training data and were used to compute initial estimates of the performance level parameters (see [10] for details). Both STAPLE and Spatial STAPLE used these initial parameter estimates using a bias value of $\kappa = 1$. For both algorithms, consensus voxels were ignored so that the excessive background did not adversely affect the segmentation accuracy and a global prior was used ($\alpha = 1$, see Eq. 11). Lastly, instead of using regional performance level estimates on a window basis, unique performance level parameters were constructed for each observed slice for Spatial STAPLE.

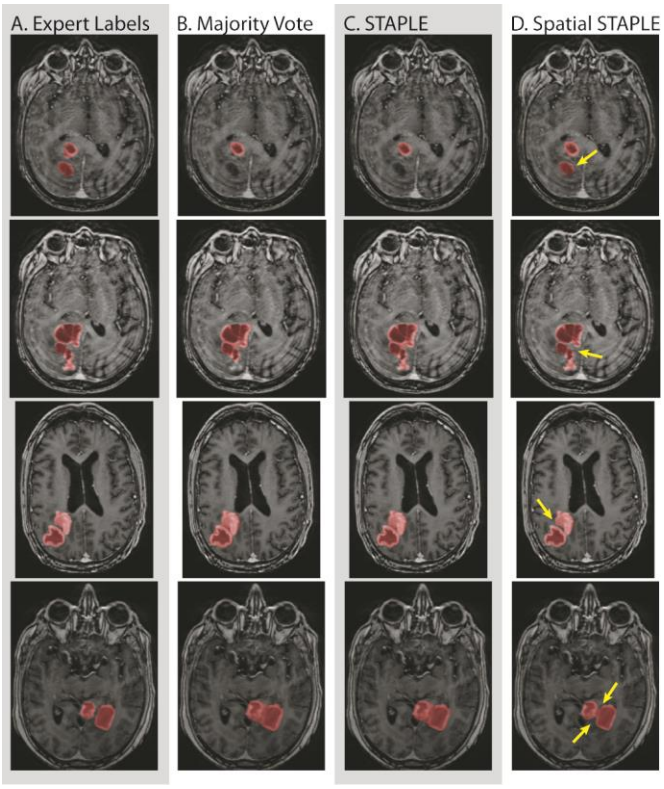


Figure III.6. Qualitative results for the human rater cancer labeling experiment. Four separate slices are shown, with the expert labels, majority vote, STAPLE and Spatial STAPLE presented for each example using 8 observations per slice. For all examples Spatial STAPLE is qualitatively superior to both majority vote and STAPLE. The arrows indicate areas of particular improvement exhibited by Spatial STAPLE.

This allows us to model the way in which raters evolve over time (e.g. a degradation over time or a learning curve).

An overall accuracy comparison between the various models of human behavior is presented in Figure III.5 with respect to increasing numbers of observations per slice (from 3 to 10). The accuracy is reported in terms of the Dice Similarity Coefficient (DSC) [140] and the boxplots represent the spread of accuracy across the 8 fused volumes. For all numbers of observations per slice Spatial STAPLE significantly outperforms STAPLE and majority vote (in a paired t-test, $p < 0.05$). Interestingly, beyond 5 observations per slice very little improvement in terms of overall accuracy is gained for any of the various algorithms. This improvement indicates that Spatial STAPLE is able to accurately characterize the spatially varying performance exhibited by the minimally trained undergraduate students. Lastly, qualitative results, using 8 observations per slice, are presented in Figure III.6. For each of the presented examples Spatial STAPLE provides a significantly more accurate estimate of the underlying segmentation than STAPLE or majority vote. Particularly, for the problem of cancer segmentation, accurately localizing cancerous regions is of the utmost of clinical importance.

3.4. Simulation of Multi-Algorithm Fusion for Whole-Brain Segmentation

Next, we evaluate the process of fusing multiple algorithms for whole brain segmentation by using a collection of 5 estimates from a locally weighted vote algorithm (using a random collection of 5 atlases per estimate). A collection of 15 whole brain segmentations were utilized with 26 labels per brain. The labels range from large structures (e.g. cerebral gray matter) to small, deep brain structures (e.g. hippocampus). Each of the brains is part of the Open Access Series of Imaging Studies (OASIS) [145] data set and the labels were acquired using the brainCOLOR protocol (http://www.braincolor.org/). A representative segmentation can be seen in Figure III.1A. The pairwise registrations were performed using an affine registration using FLIRT [98]. Since this experiment involves fusing the results of a labeling approach (the output of a N=5 locally weighted vote), only fusion techniques that don't utilize intensity differences are considered: majority vote, STAPLE and Spatial STAPLE. For both Spatial

STAPLE and STAPLE a spatially varying prior was used (i.e., $\alpha = 1$, see Eq. 19) and consensus voxels were ignored.

The results from a leave-one-out cross-validation (LOOCV) study for all 15 of the brains of interest are presented in Figure III.7. The accuracy of majority vote, STAPLE and Spatial STAPLE are considered for each of the labels individually. The boxplots represent the spread of results across the various atlases. The results indicate that Spatial STAPLE significantly outperforms (paired t-test, $p < 0.05$) both STAPLE and majority vote for nearly all of the considered labels. The only exceptions are for the left amygdala, left pallidum and the left putamen, where the results are statistically indistinguishable. Note that, particularly for the small labels, the accuracy of the estimates is less than what could be achieved with non-rigid registration and label fusion.

3.5. Empirical Experiments using Expert-labeled Head and Neck CT scans

Lastly, we analyze the accuracy of statistical fusion algorithms on an empirical multi-atlas based study. Computed tomography (CT) images were acquired from 15 patients who underwent intensity-modulated radiation therapy (IMRT) for larynx and base of tongue cancers and were expertly labeled by

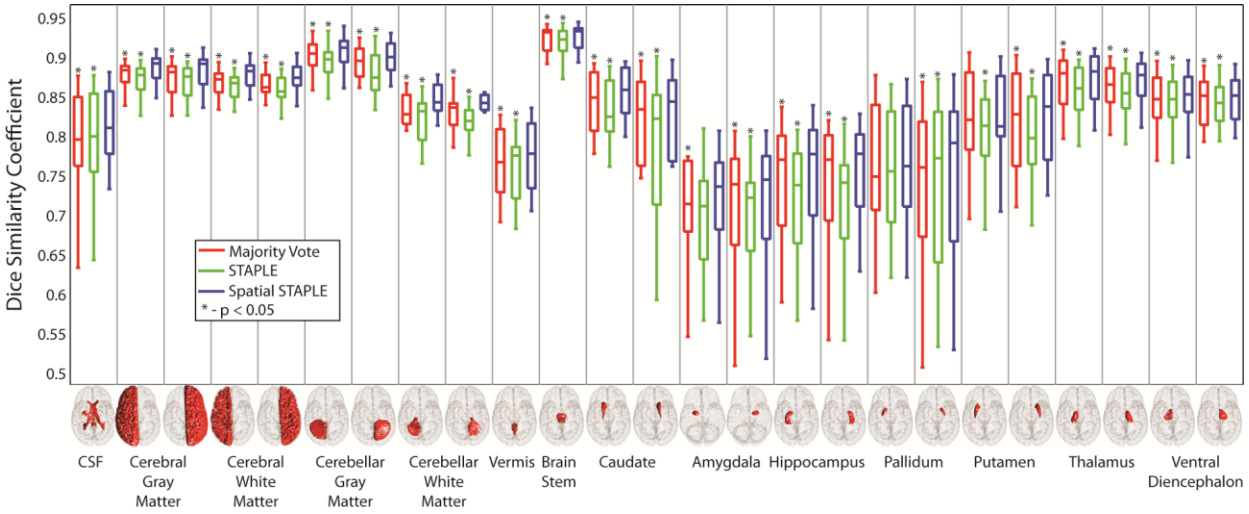


Figure III.7. Quantitative results for the simulation of meta-analysis fusion for whole brain segmentation. The presented results represent the accuracy of majority vote, STAPLE and Spatial STAPLE for all 26 labels across the 15 atlases considered in this experiment. Spatial STAPLE significantly outperforms the other algorithms for nearly all labels (excluding the left amygdala, pallidum and putamen).

an interventional radiologist. For details see [72]. Briefly, each data set has in-plane resolution of ~1mm and a slice thickness of 3mm (acquired on a Philips Brilliance Big Bore CT scanner with injection with 80mL of Optiray 320, a 68% iversol-based nonionic contrast agent). Each volume contained four segmented structures: left parotid, right parotid, right lymph node regions and thyroid. Note that 15 atlases are fairly meager for many applications, but this represents a situation where the accuracy and limits of fusion algorithms are truly tested. All analyses were performed on the full 3D volume. Following an initial affine registration to a common template, the atlases were registered using the Vectorized Adaptive Bases Registration Algorithm (VABRA) [38] and cropped to isolate the neck (~170×100×80 voxels).

Here, we compare the results between majority vote, locally weighted vote, STAPLE and Spatial STAPLE. Spatial STAPLE was used with a 0.2 window size fraction and an overlap ratio of 0.5. For both Spatial STAPLE and STAPLE a spatially varying prior was used (i.e., $\alpha = 1$, see Eq. 11) and consensus voxels were ignored. The global performance estimate utilized was from a majority vote estimate (ignoring consensus voxels) with a value of $\kappa = 1$. All accuracy comparisons were performed using the DSC. We analyze the overall accuracy across all labels as well as for each of the individual labels (i.e. the

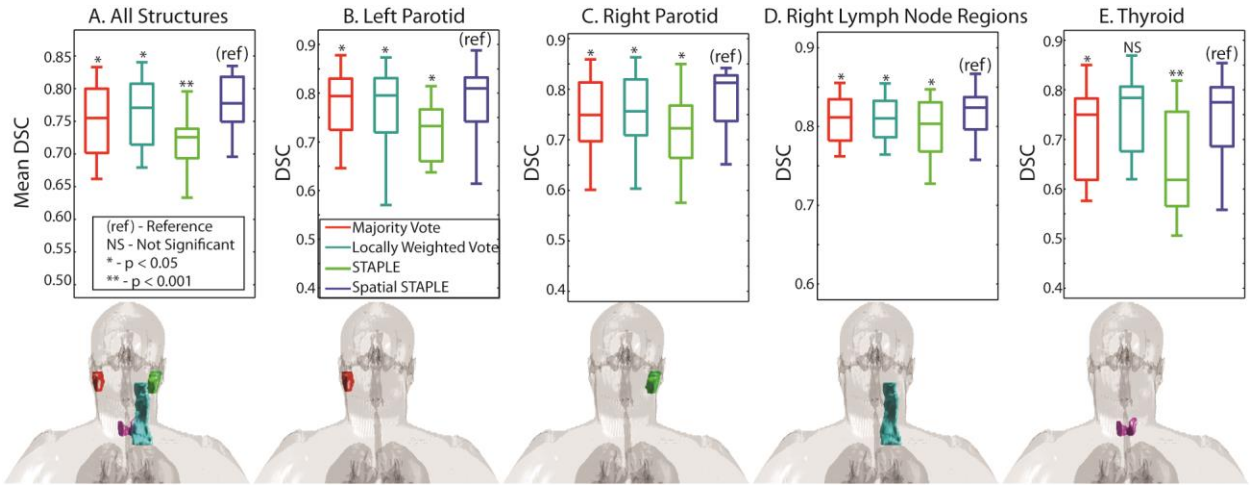


Figure III.8. Quantitative results for the segmented CT head and neck data. The mean DSC for all structures can be seen in (A). The DSC value for each of the individual algorithms can be seen in (B)-(E). Spatial STAPLE statistically outperforms locally weighted vote for all labels other than the thyroid despite the fact that Spatial STAPLE does not utilize intensity information.

left and right parotid, the right lymph node regions, and the thyroid) using a (LOOCV) study.

The accuracy of majority vote, locally weighted vote, STAPLE and Spatial STAPLE were computed for each of the 15 LOOCV iterations and plotted in Figure III.8. A paired two-sided t-test was performed to evaluate significance between the observed DSC of each approach and that of Spatial STAPLE. Spatial STAPLE results in significantly higher DSC than the other algorithms for the mean DSC and all structures except for the thyroid (where it is statistically indistinguishable from a locally weighted vote). Furthermore Spatial STAPLE is significantly superior to majority vote and STAPLE for all structures. Note, unlike a locally weighted vote, Spatial STAPLE does not use intensity information when estimating the underlying segmentation.

It is important to note that in this scenario (i.e., consensus voxels are ignored and each of the labels is primarily separated from one another), the STAPLE result is nearly equivalent to a structure-wise fusion approach (i.e. fusing each of the labels separately). Thus, the improvement with Spatial STAPLE show that the biasing prior, small window sizes (i.e. smaller than the individual structures) and large overlap provide important accuracy benefits over the traditional multi-label and structure-wise STAPLE approaches.

On average, Spatial STAPLE improves upon locally weighted vote by a DSC value of 0.01. Thus, it is important to inspect whether or not this improvement is of qualitative relevance. Representative expert labels can be seen in Figure III.9A and estimates from each of the algorithms can be seen in Figure III.9B-E. Visual inspection shows that improvements provided by Spatial STAPLE result in superior label correspondence to anatomy — particularly in the right lymph node regions. Note that defining whether or

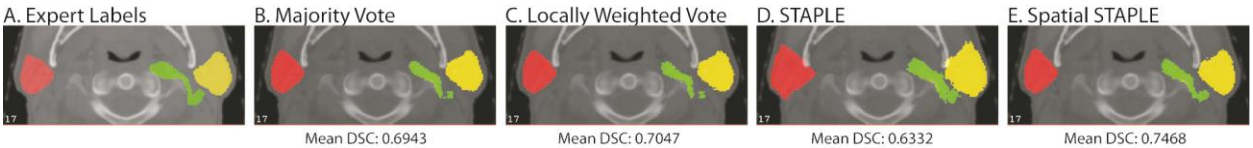


Figure III.9. Qualitative results for the segmented CT head and neck data. The average mean DSC improvement exhibited by Spatial STAPLE was approximately 0.01 DSC (Fig. 8A). Thus, it is important to assess whether or not this improvement is qualitatively visible. The truth labels can be seen in (A), with the corresponding majority vote, locally weighted vote, STAPLE and Spatial STAPLE estimates seen in (B) – (E).

not this is clinical relevance depends highly on the clinical application. Nevertheless, Spatial STAPLE significantly outperforms a locally weighted vote without using intensity information outlines the importance of regional performance level estimates in the field of label fusion.

4. Discussion and Conclusion

Herein, we derive and present Spatial STAPLE — a new algorithm for statistically fusing rater label information using a spatially varying model of rater behavior. Spatial STAPLE i) provides significant improvement over the premier label fusion techniques, ii) more accurately reflects the way in which raters and atlases make mistakes than traditional global performance metrics and iii) provides a unified framework that can be used for the gamut of label fusion applications (i.e. the fusion of human raters, multi-atlas applications and the fusion of multiple algorithms). Additionally, Spatial STAPLE is not particularly sensitive to model parameters (Figure III.4) which indicate a stable theoretical underpinning.

Like other statistical fusion algorithms [8, 50], Spatial STAPLE utilizes an E-M based approach in which the algorithm simultaneously estimates rater performance and the underlying segmentation. However, unlike its predecessors, Spatial STAPLE explicitly estimates the spatially varying performance of a collection of raters by estimating a voxelwise *performance level field* for each rater. The traditional model of rater behavior utilizes a single global confusion matrix that is utilized at all voxels. However, for many applications, global performance level parameters may fail to model the complexities of the observed labels (Figures III.5-9). By introducing a smooth, spatially-varying performance level field, the inherent errors in registration, human performance and algorithmic performance are implicitly modeled. Thus, we dramatically relax the constricting assumptions of the typical rater models and allow for significantly more freedom when estimating the models by which raters make mistakes.

Perhaps surprisingly, for the head and neck CT application, Spatial STAPLE is able to outperform a locally weighted vote in a multi-atlas context despite the fact that it is inherently a model of human behavior. This is due to several factors including 1) the registrations of the data are actually quite

poor, 2) the intensity profile for neighboring tissues are extremely similar, rendering intensity based segmentation difficult, and 3) the atlases exhibit largely spatially varying behavior due to the highly varying anatomy in the head and neck regions. While Spatial STAPLE cannot compete with intensity based techniques for many applications (e.g. whole brain segmentation), there exists a large collection of applications in a multi-atlas context for which intensity information is of limited use.

An important result presented in this paper is the ability of Spatial STAPLE to accurately estimate segmentation in the full gamut of label fusion applications. Previously, the accuracy of label fusion techniques has been highly dependent upon the application. For example, when fusing human raters, STAPLE was generally considered the best algorithm, while for other scenarios; the voting based algorithms generally provided more accurate estimations. By formulating a framework in which spatially varying behavior is characterized and semi-local performance captured, Spatial STAPLE has been shown to provide a robust and consistent framework for the fusion of human raters, multiple algorithms as well as multi-atlas fusion. Additionally, Spatial STAPLE is very versatile by modifying the way in which the semi-local performance is captured. For example, in the cancer segmentation example (Figures III.5 and 6) the semi-local performance was considered on a slice-by-slice basis, while for the multi-atlas and algorithm fusion problems the semi-local performance was modeled on a small window basis (Figures III.7-9). Given this versatility and robustness, we believe that Spatial STAPLE has the properties to cement itself as an algorithm to be considered regardless of the fusion application.

The implementation of Spatial STAPLE presented in this paper extends the concept of performance level parameters to the concept of a *performance level field* that it is varying on a voxel (local) level through analysis of regional windows and interpolation. In general there are 3 parameters that determine the manner in which the local label fusion is performed 1) the number of windows, 2) the size of the windows and 3) the type of interpolation between windows. Unfortunately, the optimal settings for these parameters depend largely on the situation. For example, in multi-atlas based segmentation we would expect the registration-driven errors to be highly local, indicating a need for many, small windows in order to accurately characterize the performance of the raters (or registered atlases). On the other hand,

for fusion of multiple minimally-trained humans who are observing labels on a slice-by-slice basis (i.e. the cancer segmentation example), the optimal parameters would allow for unique performance level parameters on each slice. This type of framework would allow the algorithm to *implicitly* capture natural human labeling phenomena such as degradation over time, or a learning curve. Spatial STAPLE provides a valuable framework for modeling the manner in which raters observe labels, which can provide optimal techniques for highly varying situations through simple parameter manipulation.

CHAPTER IV

FORMULATING IMPERFECT CORRESPONDENCE

1. Overview

Regardless of the approach, label fusion models have consistently made an implicit assumption that the use of multiple atlases results in a voxelwise, collectively unbiased representation of the target. This assumption is manifested through the fact that nearly all fusion algorithms determine the optimal label using only *directly corresponding* intensity and label information. Ergo, multi-atlas methods are generally dependent upon highly accurate registration and the use of large numbers of atlases. We are left with several problems in multi-atlas segmentation: (1) a dependence on large-scale, high-quality registrations, (2) voting-based algorithms lack the theoretical underpinning of statistical fusion observation models and (3) statistical fusion algorithms fail to incorporate intensity information. Thus, previous approaches have failed to accurately model the stochastic process of registered atlas observation error.

Meanwhile, a relatively new framework in the field of image analysis, non-local means, has gained momentum in terms of quantifying complex image characteristics (e.g., noise structure, spatially varying correspondence). In non-local means, images are deconstructed into a collection of small volumetric patches and the similarity or correspondence between these patches is quantified to learn the underlying image structure [122]. The non-local means framework has emerged in the context of image de-noising [122-127]. However, more recent work has demonstrated the applicability of non-local means to new applications such as synthesizing image contrast [146], in-painting [147], and image segmentation [56, 148].

Herein, we propose a novel statistical fusion algorithm (Non-Local STAPLE – NLS) that reformulates the STAPLE framework from a non-local means perspective. NLS models the registered atlases as collections of volumetric patches containing both intensity and label information and uses the non-local criteria [56, 122] to resolve imperfect correspondence. Through this reformulation, we seamlessly integrate exogenous intensity information into the estimation process to provide a theoretically consistent model of multi-atlas observation error. NLS provides a model in which we learn which label each atlas *would have observed* given perfect correspondence with the target. This presentation is an extension and generalization of a recently published conference paper [51]. Herein, we provide additional

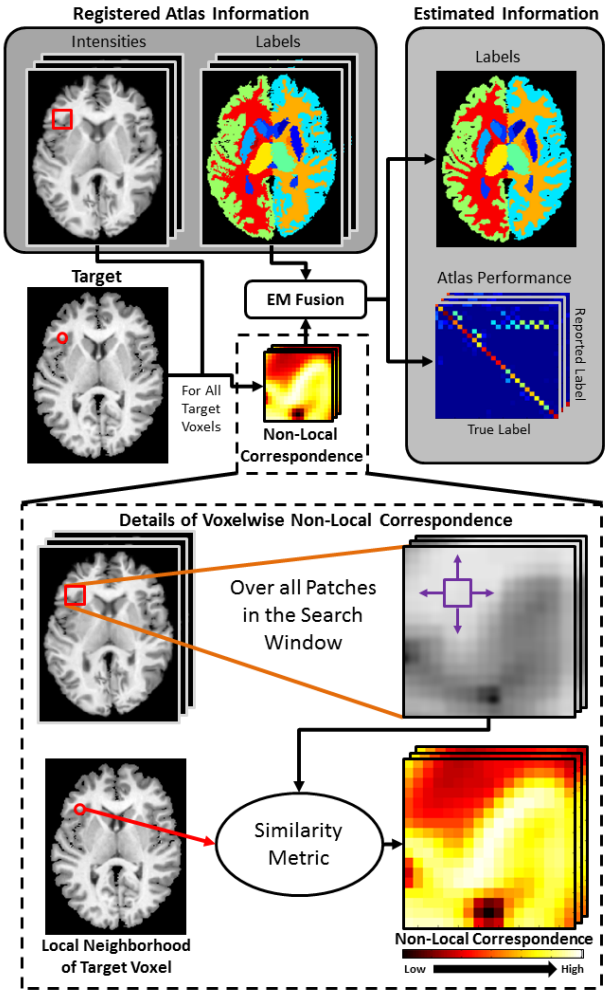


Figure IV.1. Flowchart of the Non-Local STAPLE (NLS) algorithm. NLS integrates a non-local correspondence model (using the atlas-target intensity relationships) into the estimation process. Point-wise correspondence is constructed in a traditional non-local means approach.

examples, derivations and insights that were not part of the original conference publication.

In this manuscript, we begin by deriving the theoretical basis and the parameters for initialization and convergence governing NLS. Next, we demonstrate significant improvement over the state-of-the-art fusion algorithms on two distinct datasets: (1) computed tomography (CT) images for thyroid segmentation and (2) structural magnetic resonance (MR) images for whole-brain segmentation. For whole-brain segmentation, we demonstrate that NLS dramatically lessens the need for large-scale and highly accurate non-rigid registration. Lastly, we provide insight into the sensitivity of NLS to the various model parameters, assess the optimality of the algorithm, and provide a comparison to a direct application of non-local voting.

2. Theory

The following presentation provides the theoretical model governing NLS in the commonly used Expectation-Maximization (EM) framework [116]. For clarity and consistency, the notation closely follows the presentation of the original STAPLE algorithm [8].

2.1. Problem Definition

Consider a target gray-level image represented as a vector, $\boldsymbol{I} \in \mathbb{R}^{N\times 1}$. Let $\boldsymbol{T} \in \boldsymbol{L}^{N\times 1}$ be the latent representation of the true target segmentation, where $\boldsymbol{L} = \{0, \dots, L-1\}$ is the set of possible labels that can be assigned to a given voxel. Consider a collection of R registered atlases with associated intensity values, $\boldsymbol{A} \in \mathbb{R}^{N\times R}$, and label decisions, $\boldsymbol{D} \in \boldsymbol{L}^{N\times R}$. Let $\boldsymbol{\theta} \in \mathbb{R}^{R\times L\times L}$ parameterize the performance level of raters (registered atlases). Each element of $\boldsymbol{\theta}$, $\theta_{js's}$, represents the probability that rater j observes label s' given that the true label is s at a given target voxel and the *corresponding* voxel on the associated atlas — i.e., $\theta_{js's} \equiv f\left(D_{i^*j} = s', \boldsymbol{A}_j \middle| T_i = s, I_i, \theta_{js's}\right)$, where i^* is the voxel on atlas j that corresponds to target voxel i. Throughout, the index variables i, i^* and i' will be used to iterate over the voxels, s and s' over the labels, and j over the registered atlases.

2.2. The Non-Local STAPLE Algorithm

As with other statistical fusion algorithms, NLS uses EM to estimate the true (latent) segmentation based on the target intensities, atlas information, and the rater performance level parameters (see Figure IV.1 for a graphical summary of NLS). In traditional EM terminology, the underlying voxelwise label probabilities represent the hidden data that we are estimating, and the performance level parameters, $\boldsymbol{\theta}$, represent the hidden model parameters that help determine the optimal solution for the target segmentation. The estimation of these parameters is accomplished by iterating between the E-step (i.e., the estimation of the voxelwise label probabilities) and the M-step (i.e., the estimation of the performance level parameters that maximize the expected value of the conditional log likelihood function). Before presenting the derivation of our EM-based approach, we define our non-local correspondence model, and an approximation of the performance level parameters that provides a technique for deriving the algorithm.

2.3. Non-Local Correspondence Model

In order to reformulate the traditional STAPLE model of rater behavior from a non-local means perspective, we need to define an appropriate non-local correspondence model. Given a voxel on the target image, i, this correspondence model provides a technique for determining the corresponding voxel on a given atlas, i^*. In our model, there are two primary components that are required to define the non-local correspondence: (1) the intensity similarity model between a given atlas voxel and the target voxel of interest, and (2) the spatial compatibility between two voxel locations in the common target image coordinate system.

First, there are several options that could be used to define the intensity similarity between a given atlas voxel and the target voxel (e.g., correlation coefficient [53], mutual information [48], Gaussian intensity difference [59]). Herein, we use a Gaussian difference model, which, assuming proper intensity normalization, has been shown to be highly successful, particularly on neurological applications [46, 59, 123].

Second, we need to define a metric for the spatial compatibility between a given atlas voxel and the target voxel in image space. Traditional non-local means algorithms for image de-noising [122-125] weight all voxels equally, regardless of the distance between the voxels in image space. However, in order to translate non-local means to segmentation-based applications, limited search regions are typically defined in order to prevent confusion between structures with similar intensity profiles [56, 148]. Here, we employ a Gaussian window-based model so that highly local voxels are more highly weighted. This reflects our desire to estimate that the underlying corresponding voxel i^* is both similar to the target voxel and, due to the registration process, generally close in terms of the target image coordinate system.

Together, we define the probability of correspondence between an atlas voxel and the given target voxel (i.e., $f\left(A_{i'j}\middle|I_i\right)$) to be the product of two Gaussian distributions.

$$f\left(A_{i'j}\middle|I_i\right) \equiv \alpha_{ji'i} = \frac{1}{Z_\alpha}\exp\left(-\frac{\left\|\wp\left(A_{i'j}\right)-\wp(I_i)\right\|_2^2}{2\sigma_i^2}\right)\exp\left(-\frac{\mathcal{E}_{ii'}^2}{2\sigma_d^2}\right) \tag{4.1}$$

where the first distribution is the intensity similarity model, the second distribution is the spatial compatibility model, and Z_α is a partition function. In the intensity similarity model, $\wp(\cdot)$ is the set of intensities in the *patch neighborhood* of a given intensity location and σ_i is the standard deviation of the assumed distribution. In the spatial compatibility model, $\mathcal{E}_{ii'}$ is the Euclidean distance between voxels i and i' in image space and σ_d is the corresponding standard deviation.

Lastly, the partition function, Z_α enforces the constraint that

$$\sum_{i'\in\mathcal{N}(i)} \alpha_{ji'i} = 1. \tag{4.2}$$

where $\mathcal{N}(i)$ is the set of voxels in the *search neighborhood* of a given target voxel. Through this constraint, $\alpha_{ji'i}$ can be directly interpreted as the probability that voxel i' on atlas j is the latent corresponding voxel, i^*, to a given target voxel i.

2.4. Approximation of the Latent Performance Level Parameters

The following derivation of NLS hinges upon knowledge of the voxel i^* on atlas j that directly corresponds to voxel i on the target image. If the directly corresponding voxel was known, then the ideal non-local correspondence model would be known and we could ignore the intensity relationships to use a typical definition of the underlying performance level parameters.

$$f\left(D_{i^*j} = s', \boldsymbol{A}_j \middle| T_i = s, I_i, \theta_{js's}\right) = f\left(D_{i^*j} = s' \middle| T_i = s, \theta_{js's}\right) \equiv \theta_{js's} \tag{4.3}$$

Unfortunately, this corresponding voxel, i^*, is unknown and we are forced to approximate it using the previously defined non-local correspondence model. Using the model in Eq. 1, we can approximate this relationship by taking the expected value of $f(D_{i^*j} = s', \boldsymbol{A}_j | T_i = s, I_i, \theta_{js's})$ across the atlas image. Using an assumption of conditional independence between the labels and intensity, we approximate the desired density function

$$\begin{aligned} f(D_{i^*j} = s', \boldsymbol{A}_j | T_i = s, I_i, \theta_{js's}) &\approx E\left[f\left(\boldsymbol{D}_j, \boldsymbol{A}_j \middle| T_i = s, I_i, \boldsymbol{\theta}_{js}\right)\right] \\ &= E\left[f\left(\boldsymbol{D}_j \middle| T_i = s, \boldsymbol{\theta}_{js}\right) f\left(\boldsymbol{A}_j \middle| I_i\right)\right] \\ &= \sum_{i' \in \mathcal{N}(i)} f\left(D_{i'j} = s' \middle| T_i = s, \theta_{js's}\right) f\left(A_{i'j} \middle| I_i\right) \\ &= \sum_{i' \in \mathcal{N}(i)} \theta_{js's} \alpha_{ji'i} \end{aligned} \tag{4.4}$$

where $\mathcal{N}(i)$ is the set of voxels in the *search neighborhood* of voxel i, and $\alpha_{ji'i}$ is the previously defined non-local correspondence model (Eq. 1).

As in [59], we assume conditional independence between the labels and intensity, which seemingly neglects their complex relationships. However, our assumption is that the information gained from inclusion of the atlas intensity is related to understanding the lack of local correspondence between the target and the atlas, which, through the estimation process, indirectly models the complex label-intensity relationships.

Additionally, it is important to note that this model of the performance level parameters is inherently an approximation based upon an assumed *a priori* distribution (Eq. 1) governing the non-local correspondence between the target and the atlases. Ideally, the non-local correspondence parameters

would be treated as additional model parameters that are iteratively updated in the M-step of the subsequent EM algorithm. Unfortunately, there are two primary limitations that prevent the construction of this type of idealized model. First, this model makes solving the M-step of the algorithm mathematically difficult as we would be forced to simultaneously estimate the raters' performance and the voxel(s) that represent the true underlying correspondence. Second, it dramatically increases the number of parameters that we would be attempting to estimate. To illustrate, given a non-local search neighborhood consisting of K voxels, the number of augmented model parameters would be approximately $K \times N \times R$ which leaves an underdetermined system given the amount of data that is available to estimate these parameters. Regardless, despite these limitations, the proposed model approximation captures many of the same benefits that would likely be achieved assuming the "ideal" approach were possible to construct.

2.5. E-Step: Estimation of the Voxelwise Label Probabilities

Let $\boldsymbol{W} \in \mathbb{R}^{L \times N}$, where $W_{si}^{(k)}$ represents the probability that the true label associated with voxel i is label s at iteration k of the algorithm given the provided information and model parameters

$$W_{si}^{(k)} \equiv f\left(T_i = s \middle| \boldsymbol{D}, \boldsymbol{A}, \boldsymbol{I}, \boldsymbol{\theta}^{(k)}\right). \tag{4.5}$$

Using a Bayesian expansion and the assumed conditional independence between the registered atlas observations, Eq. 5 can be re-written as

$$W_{si}^{(k)} = \frac{f(T_i = s) \prod_j f(D_{i^*j} = s', \boldsymbol{A}_j | T_i = s, I_i, \theta_{js's}^{(k)})}{\sum_n f(T_i = n) \prod_j f(D_{i^*j} = s', \boldsymbol{A}_j | T_i = n, I_i, \theta_{js'n}^{(k)})} \tag{4.6}$$

where $f(T_i = s)$ is a voxelwise *a priori* distribution of the underlying segmentation, and D_{i^*j} is the label decision by atlas j at the atlas image voxel i^* that corresponds to voxel i on the target image. Note that the denominator of Eq. 6 is simply the solution for the partition function that enables $\boldsymbol{W}$ to be a valid probability mass function (i.e., $\sum_s W_{si} = 1$).

As previously noted, we do not know the corresponding atlas voxel. Thus, using the non-local correspondence model (Eq. 1) and the provided approximation (Eq. 4), we can approximate the final solution for the voxelwise label probabilities

$$W_{si}^{(k)} = \frac{f(T_i = s)\prod_j \sum_{i' \in \mathcal{N}(i)} \theta_{js's}^{(k)} \alpha_{ji'i}}{\sum_n f(T_i = n)\prod_j \sum_{i' \in \mathcal{N}(i)} \theta_{js'n}^{(k)} \alpha_{ji'i}}. \tag{4.7}$$

where, it is assumed that $D_{i'j} = s'$.

2.6. M-Step: Estimation of the Performance Level Parameters

The estimate of the performance level parameters (M-step) is obtained by finding the parameters that maximize the expected value of the conditional log likelihood function (i.e., using the result in Eq. 7).

$$\begin{aligned} \boldsymbol{\theta}_j^{(k+1)} &= \arg\max_{\boldsymbol{\theta}_j} \sum_i E\left[\ln f(D_{i^*j} = s', \boldsymbol{A}_j | T_i = s, I_i, \theta_{js's}^{(k)}) \,\middle|\, \boldsymbol{D}, \boldsymbol{A}, \boldsymbol{I}, \boldsymbol{\theta}^{(k)}\right] \\ &= \arg\max_{\boldsymbol{\theta}_j} \sum_i \sum_s W_{si}^{(k)} \ln f(D_{i^*j} = s', \boldsymbol{A}_j | T_i = s, I_i, \boldsymbol{\theta}_j^{(k)}) \end{aligned} \tag{4.8}$$

Noting the constraint that each row of the rater performance level parameters must sum to unity to be a valid probability mass function (i.e., $\sum_{s'} \theta_{js's}^{(k)} = 1$), we can maximize the performance level parameters for each element by using a Lagrange Multiplier (λ) [117] to formulate the constrained optimization problem. Following this procedure, we obtain

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta_{js's}} \left[\sum_i \sum_s W_{si}^{(k)} \ln\left(f\left(D_{i^*j} = s', \boldsymbol{A}_j \middle| T_i = s, I_i, \boldsymbol{\theta}_j^{(k)}\right)\right) + \lambda \sum_{s'} \theta_{js's}\right] \\ 0 &= \sum_i W_{si}^{(k)} \frac{\frac{\partial}{\partial \theta_{js's}}\left[f\left(D_{i^*j} = s', \boldsymbol{A}_j \middle| T_i = s, I_i, \boldsymbol{\theta}_j^{(k)}\right)\right]}{f\left(D_{i^*j} = s', \boldsymbol{A}_j \middle| T_i = s, I_i, \boldsymbol{\theta}_j^{(k)}\right)} + \lambda. \end{aligned} \tag{4.9}$$

However, in order to solve for $\theta_{js's}$ we have to utilize the approximation presented in Eq. 4. The density function of interest, $f\left(D_{i^*j} = s', \boldsymbol{A}_j \middle| T_i = s, I_i, \boldsymbol{\theta}_j^{(k)}\right)$, appears in both the numerator and the denominator. In the denominator, we see the exact density function that we are trying to maximize; thus, we substitute the direct definition of the performance level parameters presented in Eq. 3. In the numerator, however, we need to take the derivative of this density function with respect to the current element of the

performance level parameters (and the dependence structure is not apparent in Eq. 3). To capture the inherent noise and lack of local correspondence between the target and the atlases, we use the approximation of this density function (i.e., Eq. 4) in the numerator. Using these substitutions and some straightforward algebraic manipulation we obtain

$$\begin{aligned} 0 &= \frac{\sum_i W_{ni}^{(k)} \frac{\partial}{\partial \theta_{js's}} \left[\sum_{i' \in \mathcal{N}(i)} \theta_{js's}^{(k)} \alpha_{ji'i} \right]}{\theta_{js's}} + \lambda \\ 0 &= \frac{\sum_i W_{ni}^{(k)} \frac{\partial}{\partial \theta_{js's}} \left[\sum_{s'} \sum_{i' \in \mathcal{N}(i): D_{i'j} = s'} \theta_{js's}^{(k)} \alpha_{ji'i} \right]}{\theta_{js's}} + \lambda \\ 0 &= \frac{\sum_i W_{ni}^{(k)} \sum_{i' \in \mathcal{N}(i): D_{i'j} = s'} \alpha_{ji'i}}{\theta_{js's}} + \lambda \\ \theta_{js's} &= \frac{\sum_i W_{ni}^{(k)} \sum_{i' \in \mathcal{N}(i): D_{i'j} = s'} \alpha_{ji'i}}{-\lambda} \end{aligned} \tag{4.10}$$

Finally, solving for the Lagrange Multiplier leaves the final solution for each element of the performance level parameters

$$\theta_{js's}^{(k+1)} = \frac{\sum_i \left(\sum_{i' \in \mathcal{N}_s(i): D_{i'j} = s'} \alpha_{ji'i} \right) W_{si}^{(k)}}{\sum_i W_{si}^{(k)}}. \tag{4.11}$$

2.7. Initialization, Model Parameters, and Detection of Convergence

As with all of the algorithms that have been presented in the STAPLE family, NLS can be initialized using either an initial estimate of the performance level parameters or the voxelwise label probabilities. For all of the presented experiments, NLS was initialized with performance parameters equal to 0.95 along the diagonal and randomly setting the off-diagonal elements to fulfill the required constraints. Note that initializing NLS in this way is essentially the same as initializing the voxelwise label probabilities to that of a majority vote.

For all presented experiments, the voxelwise label prior, $f(T_i = s)$, was initialized using the label probabilities from a "weak" log-odds majority vote (i.e., decay coefficient set to 0.5 voxels) [59]. We found that initializing in this manner provided enough spatial information for NLS to consistently converge to a desired optimum, without being too spatially restrictive. Alternative approaches could be to

(1) initialize using a global prior (i.e., the same probabilities for every voxel), or (2) use the output of another segmentation algorithm.

There are several parameters in the non-local correspondence model that need to be set in order to efficiently utilize NLS. First, there are two neighborhood parameters that need to be initialized: the search neighborhood, $\mathcal{N}(i)$, and the patch neighborhood, $\wp(\cdot)$. Both of these parameters are functions of the input data (e.g., the resolution of the images, the quality of registration). For all of the presented experiments we used a search neighborhood of size $11 \times 11 \times 11$ voxels centered at the target voxel of interest. We found that inter-subject registrations were of a high enough quality that a search neighborhood of this size was able to consistently capture the underlying non-local correspondence. For the patch neighborhood, several potential sizes are considered (all of which are centered at the voxel of interest) and the benefits and detriments of varying this value are discussed later in this manuscript. The two standard deviation parameters that need to be set are σ_i and σ_d, which control the impact of the intensity difference and the Euclidean distance-based decay, respectively. In general, σ_i is a function of the intensity normalization process and, thus, spread of intensity values. The parameter σ_d can be thought of as a proxy for the search neighborhood. Unless otherwise noted, these values were set to 0.1 and 2, for σ_i and σ_d, respectively. These "default" values were obtained during the coding implementation of the proposed algorithm and were tested on a single whole-brain volume in order to obtain reasonable results. Note that this is a non-ideal approach for determining these parameters as it (1) slightly biases the presentation of results, and (2) does not guarantee the optimality of the parameters (as indicated in Figure IV.8). For future applications, where distinct and independent testing and training data are available, it would be more appropriate to determine the optimal parameter values using the training data only (i.e., the available atlases).

Convergence of NLS was detected by monitoring the change in the performance level parameters between consecutive iterations. As with the original STAPLE algorithm, we considered the algorithm to have converged when the average change in the on-diagonal elements of the performance level

parameters fell below 10^{-4}. For all presented experiments, convergence occurred in fewer than 10 iterations.

Lastly, while not necessarily a model parameter, "consensus voxels" (i.e., voxels where all raters agree) were ignored during the estimation process. Due to the non-local nature of the algorithm, consensus voxels were determined in two subsequent steps. First, an initial "consensus voxels" estimate was obtained by finding all voxels for which $\max_s f(T_i = s) > 0.95$. Second, this initial estimate was post-processed to include a safety margin around the estimated non-consensus voxels that is defined by the search neighborhood (i.e., all voxels within the search neighborhood of a non-consensus voxel were determined to be non-consensus as well). This accomplishes two tasks: (1) it improves the runtime of the algorithm and (2) it prevents the performance level parameters from being unnecessarily biased due to the inclusion of highly "consensus" regions [9, 50].

3. Methods and Results

An implementation of the Non-Local STAPLE algorithm is available as part of the Java Image Science Toolkit (JIST, www.nitrc.org/projects/jist).

3.1. Baseline Algorithms

Our first baseline algorithm is a log-odds majority vote (MV) [59]. For all presented experiments the decay coefficient was set to unity, as suggested in [59]. Our second baseline is a locally weighted vote (LWV) [48, 57, 59]. LWV procedures have come to represent the state-of-the-art fusion strategy as they provide consistent improvement over both MV and globally-weighted approaches. The implementation presented here is the same as suggested in [59]. Note that a LWV has a parameter that is essentially identical to the σ_i parameter in NLS (see Eq. 1). For fairness of comparison, this parameter was initialized to the same value (herein, 0.1) for both algorithms. Our next baseline is the original STAPLE algorithm [8]. Due to the amount of overlap between STAPLE and NLS the same parameter values were used when applicable. Thus, the algorithms were equivalently initialized, the same values

were used for the voxelwise label prior, $f(T_i = s)$, "consensus voxels" were ignored using the same discriminant criteria, and convergence was detected using the same threshold.

Our last baseline algorithm is Spatial STAPLE [49, 52, 54]. Spatial STAPLE represents an extension to the traditional STAPLE framework that allows for the estimation of a smooth spatially-varying performance level field instead of global performance level parameters and has been shown to provide robust and accurate multi-atlas segmentations. Where applicable, Spatial STAPLE was utilized using identical parameters to NLS and STAPLE. In addition, the performance level parameters were calculated on a voxelwise basis using a half-window size of 10mm in all cardinal directions. Note that Spatial STAPLE is very similar to another recently proposed algorithm – Local STAPLE MAP [54]. The primary difference is the way in which the performance level parameters are kept stable. Here, Spatial STAPLE uses a non-parametric distribution governed by an initial estimate from the original STAPLE algorithm as opposed to the parametric beta distribution that is proposed in Local STAPLE MAP. Investigation into the optimal way to maintain performance level stability is outside of the scope of this manuscript.

3.2. Motivating Simulation

Before presenting the empirical results, we present a toy simulation to demonstrate the limitations of the traditional STAPLE model of rater behavior (Figure IV.2). A single 2D slice (144 x 191 voxels) from a manually labeled whole-brain dataset was used as the basis for the presented simulation models (see the "Empirical Evaluation" section for details on the dataset). The presented slice has 4 non-background labels (left/right cerebral gray matter and left/right cerebral white matter) and the accuracy of the presented algorithms is presented in terms of the mean Dice Similarity Coefficient [140] across these labels. For each presented example, 8 label observations were simulated per fusion estimate. In Figure IV.2 we present three different models of rater observation behavior:

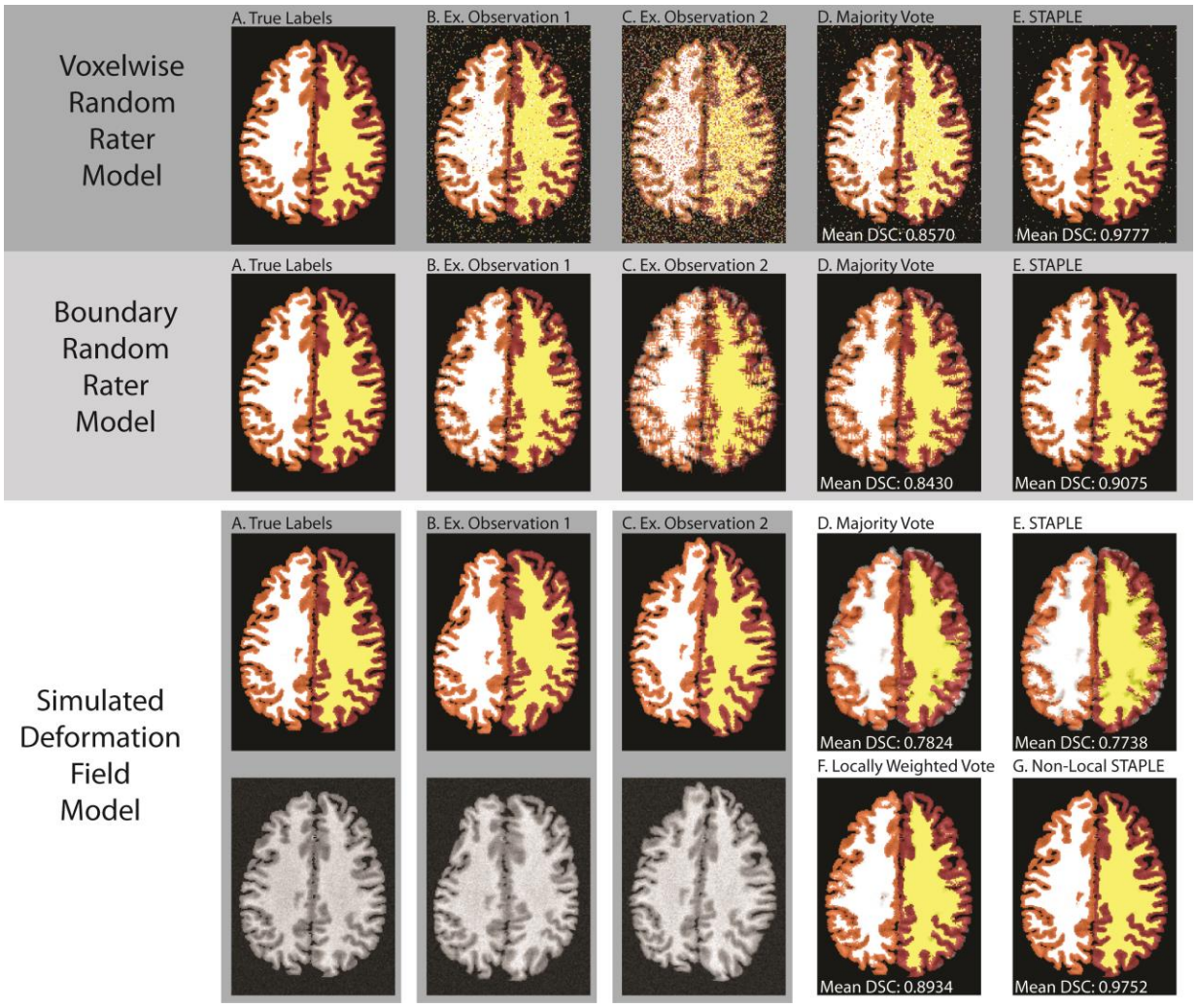


Figure IV.2. Simulated models of rater behavior and their impact on fusion performance. The first two examples present traditional models of human observation behavior, and, for both models, STAPLE substantially outperforms a majority voting based approach. In contrast, the third example simulates a typical multi-atlas observation model. In this case, STAPLE is outperformed by a majority vote. Additionally, the multi-atlas fusion approaches that utilize the target-atlas intensity relationships (e.g., locally weighted vote and the proposed Non-Local STAPLE) provide substantial improvement.

- The first observation model represents a "voxelwise random rater model" [8, 11, 50] in which simulated confusion matrices are constructed for each rater. Simulated observations are generated through Monte Carlo sampling of these confusion matrices given the true segmentation. Here, confusion matrices were randomly constructed with constant on-diagonal values linearly distributed between 0.5 and 0.9.

- The second observation model represents a "boundary random rater model" [10, 11, 49, 50] in which the boundary voxels on the true segmentation are randomly shifted. The shift amount was randomly sampled from a zero-mean Gaussian distribution that is unique to each rater. The standard deviation values of these distributions were linearly distributed between 0.5 and 2.
- The last observation model represents a "simulated deformation field model" in which simulated deformation fields are applied to the true labels by sampling a sixth-order Chebyshev polynomial with random coefficients unique to each rater. These coefficients were randomly sampled from a zero-mean Gaussian distribution with standard deviation equal to unity.

The first two examples are typical simulated models of human rater observation behavior, and, in both cases, STAPLE provides substantial improvement over a MV. To contrast, the third example simulates a typical multi-atlas observation model, in which random deformations are applied to a target image. In this case, STAPLE is slightly outperformed by a MV, which highlights the lack of applicability of STAPLE's observations model to a multi-atlas context. Additionally, using the intensity images of the simulated "atlases" in the third simulation model, we show that a LWV and NLS provide substantial improvement over the traditional "human rater" fusion models (i.e., MV and STAPLE) that ignore the target-atlas intensity relationships.

3.3. Empirical Evaluation

We consider two distinct empirical datasets. Our first dataset is a collection of 15 CT head and neck atlases used for thyroid segmentation. The images used in this experiment were collected from consenting patients who underwent intensity-modulated radiation therapy. The patients were injected with 80mL of Optiray 320, a 68% iversol-based nonionic contrast agent. Each image has a voxel size of $1 \times 1 \times 3\ \mathrm{mm}^3$. The expert labels were obtained from a local expert radiologist and verified by multiple experienced human raters. Note that 5 of the 15 patients in this data set underwent a surgical procedure that split their thyroid into two distinct sections.

Our second dataset is a collection of 15 Magnetic Resonance (MR) images of the brain as part of the Open Access Series of Imaging Studies (OASIS) [145] dataset. This data was expertly labeled courtesy of Neuromorphometrics, Inc. (Somerville, MA) and provided under a non-disclosure agreement. A refined dataset (using the OASIS brains and a subtly revised labeling protocol) has recently been made available as part of the MICCAI 2012 workshop on multi-atlas labeling. This data is available at the following URL: https://masi.vuse.vanderbilt.edu/workshop2012/ or directly from Neuromorphometrics. For each atlas, a collection of 26 labels (including background) were considered: ranging from large structures (e.g., cortical gray matter) to smaller deep brain structures (see Figure IV.5 for a list of all labels). Note that all of the cortical surface labels were combined to form left and right cortical gray matter labels. All images are 1mm isotropic resolution and, for ease of analysis, the brain region was extracted.

Note that, while all baseline algorithms were considered, the STAPLE results are not shown for the whole-brain segmentation problem as it has been demonstrated to be consistently outperformed by a LWV for whole-brain segmentation [48, 49, 57, 59]. Nevertheless, the MV results are shown in order to provide a reference baseline for registration performance and segmentation accuracy.

3.4. Pre-Processing and Analysis

All pairwise registrations were performed using an initial affine registration [98], and, when noted, all pairwise non-rigid registrations were performed using the Vectorized Adaptive Bases Registration Algorithm (VABRA) [38]. After registration, the images were (1) cropped so that excess background was removed, and (2) intensity normalized such that the 25^{th} and 75^{th} percentiles of the range of the non-background intensity values were set to 0 and 1, respectively. Quantitative accuracy was assessed using the Dice Similarity Coefficient (DSC) [140], Hausdorff distance [149], and mean surface distance. The surface distance metrics were computed unidirectionally in terms of the distance from the expert labels to the estimated segmentation.

3.5. Thyroid Segmentation Results

Our first experiment analyzes the fusion accuracy for segmentation of the thyroid. In addition to the benchmarks, NLS was run using various patch neighborhood, $\wp(\cdot)$, sizes ($1 \times 1 \times 1$, $3 \times 3 \times 3$, $5 \times 5 \times 3$, and $7 \times 7 \times 3$ voxels), all of which were centered at the voxel of interest. Due to the slice thickness of 3mm, the third dimension of the patch neighborhoods were not increased beyond 3 voxels. We performed a leave-one-out cross-validation experiment (i.e., 14 atlases per segmentation estimate) to assess fusion accuracy. The results of this experiment are presented in Figure IV.3.

The quantitative results, in terms of the spread across the 15 atlases, can be seen in Figure IV.3A.

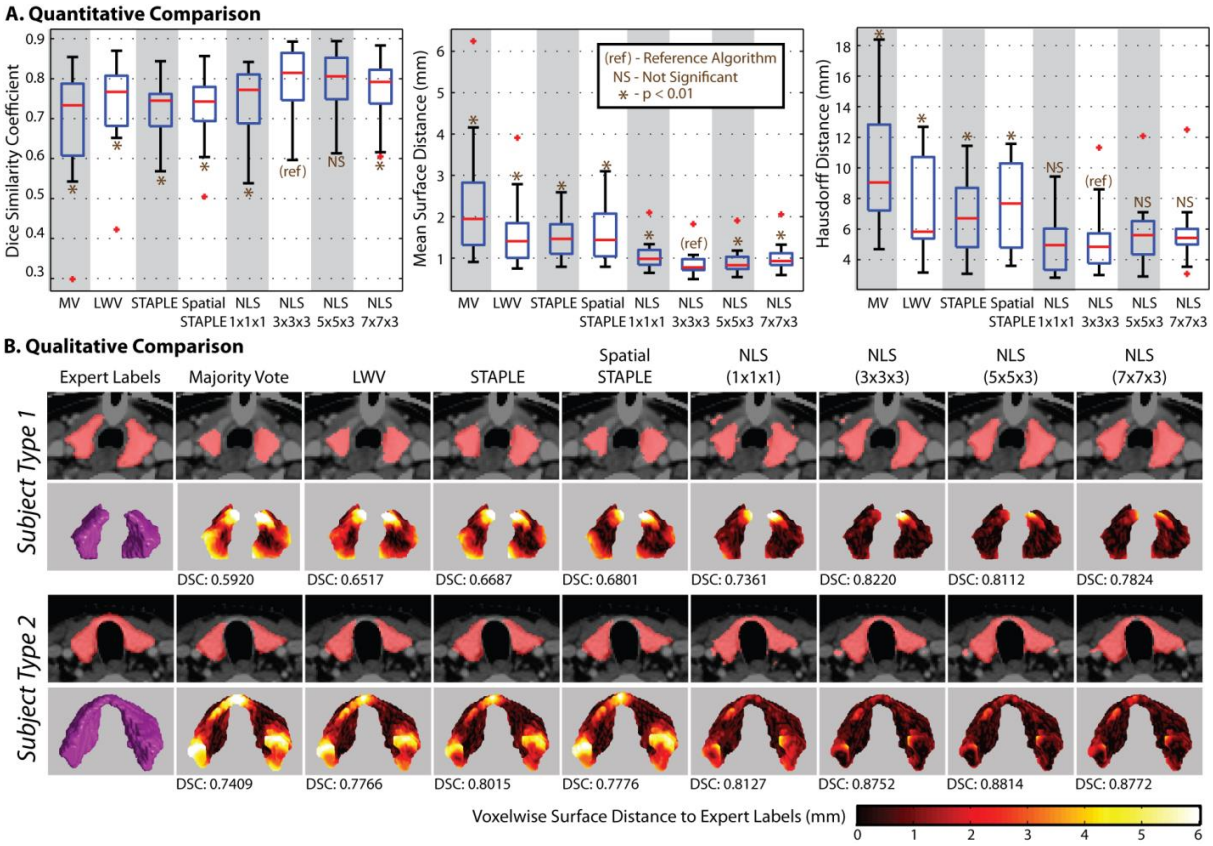


Figure IV.3. Results of the empirical multi-atlas segmentation of the thyroid. The quantitative results (A) show that NLS provides significant improvement in terms of the DSC, Hausdorff distance, and mean surface distance, with a $3 \times 3 \times 3$ patch neighborhood as the most consistent performer. The qualitative results (B) support the quantitative improvement and demonstrate that NLS provides substantial improvement in shape, boundary, and point-wise surface distance error. Note that "Subject Type 1" underwent a surgery to surgically bisect the thyroid.

The NLS based approaches provide significant improvement ($p < 0.01$, paired t-test) over all of the considered baseline algorithms in terms of the DSC, Hausdorff distance and mean surface distance. NLS using a $3 \times 3 \times 3$ (voxels) patch neighborhood size was the most consistent performer as it significantly outperformed ($p < 0.01$, paired t-test) the other NLS based approaches in terms of the DSC and the mean surface distance. The median DSC performance was improved by 0.05 over a LWV and 0.08 over STAPLE. Only the NLS based approaches achieved submillimetric accuracy in terms of the mean surface distance between the expert labels and the segmentation estimates. Additionally, NLS using a $3 \times 3 \times 3$ (voxels) patch neighborhood provided over 1mm improvement over a LWV and over 2mm improvement over STAPLE and Spatial STAPLE in terms of the Hausdorff distance.

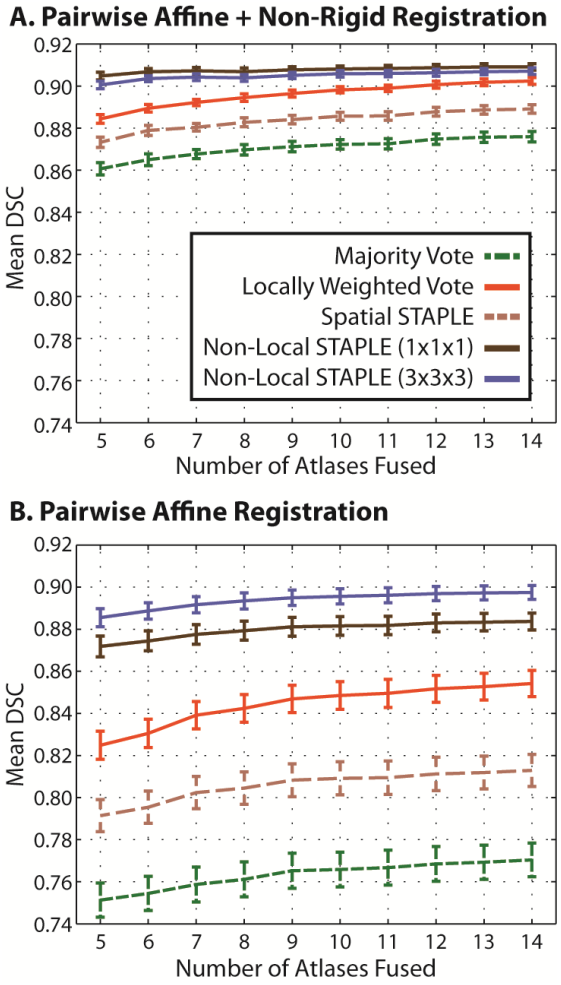


Figure IV.4. Overall accuracy, in terms of mean DSC, comparison for whole-brain segmentation. For both pairwise non-rigid and pairwise affine registration procedures, NLS provides significant improvement over traditional fusion approaches.

Qualitative results are presented in Figure IV.3B, where, for all considered algorithms a representative slice and a 3D rendering of the point-wise surface distance error is presented. Example results are presented for a representative patient that underwent a surgery to bisect the thyroid (subject type 1) and a representative subject that did not (subject type 2). The segmentations from NLS are all qualitatively superior to the other baseline algorithms as they more accurately estimated the underlying shape and size and resulted in substantial reductions in point-wise surface distance error. For small patch neighborhoods (e.g., $1 \times 1 \times 1$ – a single voxel) it is evident that high quality boundaries are estimated, but "speckle noise" is more likely to be apparent. Alternatively, for larger windows, estimations are smoother but sacrifice the high quality boundary estimation. Note that only the NLS based approaches correctly estimated the connected topography of the second subject.

3.6. Whole-Brain Segmentation Results

Our second experiment analyzes fusion accuracy for whole-brain segmentation. For this experiment, NLS was run using both $1 \times 1 \times 1$ (voxel) and $3 \times 3 \times 3$ (voxel) patch neighborhoods. The results of this experiment are presented using a pairwise non-rigid registration procedure and a pairwise affine registration procedure. For both registration procedures, the overall accuracy (in terms of mean

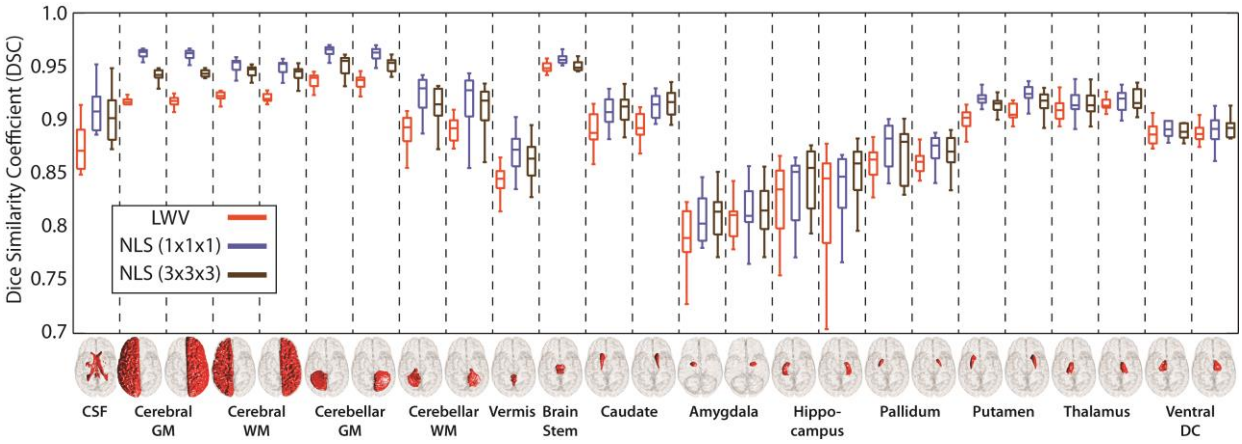


Figure IV.5. Per-label accuracy comparison on the whole-brain segmentation problem using a pairwise non-rigid registration procedure. NLS provides consistent improvement over locally weighted voting. In this case, NLS using a single voxel patch neighborhood consistently outperformed a larger ($3 \times 3 \times 3$) patch neighborhood.

DSC) was assessed using a cross-validation experiment with between 5 and 14 atlases per target. Additionally, the per-label accuracy was assessed using 5 atlases per target.

The results of the overall accuracy comparison for both registration procedures are summarized in Figure IV.4. The results indicate that, for both the pairwise non-rigid registration (Figure IV.4A) and the pairwise affine registration (Figure IV.4B), NLS demonstrates significant improvement ($p < 0.001$, paired t-test) over MV, LWV and Spatial STAPLE regardless of the number of atlases fused. For the non-rigid registration, NLS using a single voxel patch neighborhood provided a small, yet consistent, improvement over the larger $3 \times 3 \times 3$ (voxels) patch neighborhood. Interestingly, the opposite was true for the affine registration, where the larger neighborhood provided consistent improvement over the single voxel neighborhood. This difference indicates the importance of using larger patch neighborhoods when the quality of registration is diminished, and the expected correspondence is highly non-local. Additionally, for both registration procedures, NLS using only 5 atlases exhibited significant improvement ($p < 0.05$) over a LWV using all 14 available atlases. Note that, unlike [54], Spatial STAPLE is consistently outperformed by a LWV. This disparity is primarily due to the fact that the structures presented here are highly dependent upon their intensity characteristics. In [54], they focus on cortical segmentation – a problem in which intensity information provides little benefit in terms of distinguishing between adjacent

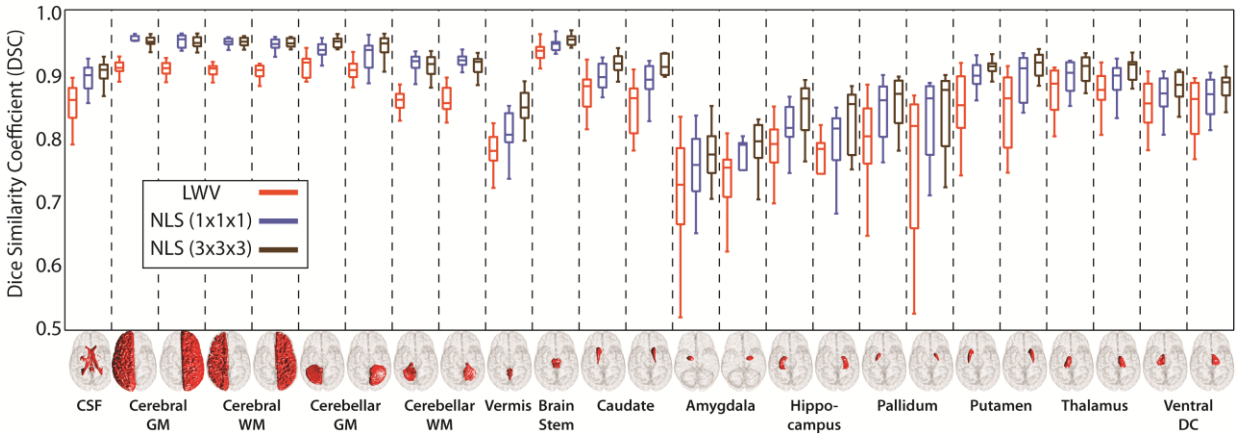


Figure IV.6. Per-label accuracy comparison on the whole-brain segmentation problem using a pairwise affine registration procedure. As in Figure IV.5, NLS provides consistent improvement over locally weighted voting. In this case, NLS using a larger ($3 \times 3 \times 3$) patch neighborhood consistently outperformed a single voxel patch neighborhood.

labels.

The per-label results for the non-rigid (Figure IV.5) and affine (Figure IV.6) registration procedures demonstrate consistent improvement over a LWV regardless of label size, location and shape. For the non-rigid results, NLS using a single voxel patch neighborhood resulted in significantly superior ($p < 0.05$) results over LWV on 23 out of 25 labels and for 16 out of 25 labels over NLS using a $3 \times 3 \times 3$ (voxels) patch neighborhood. For the affine results, NLS using a $3 \times 3 \times 3$ (voxels) patch neighborhood resulted in significant improvement ($p < 0.05$) over LWV on all considered labels and for

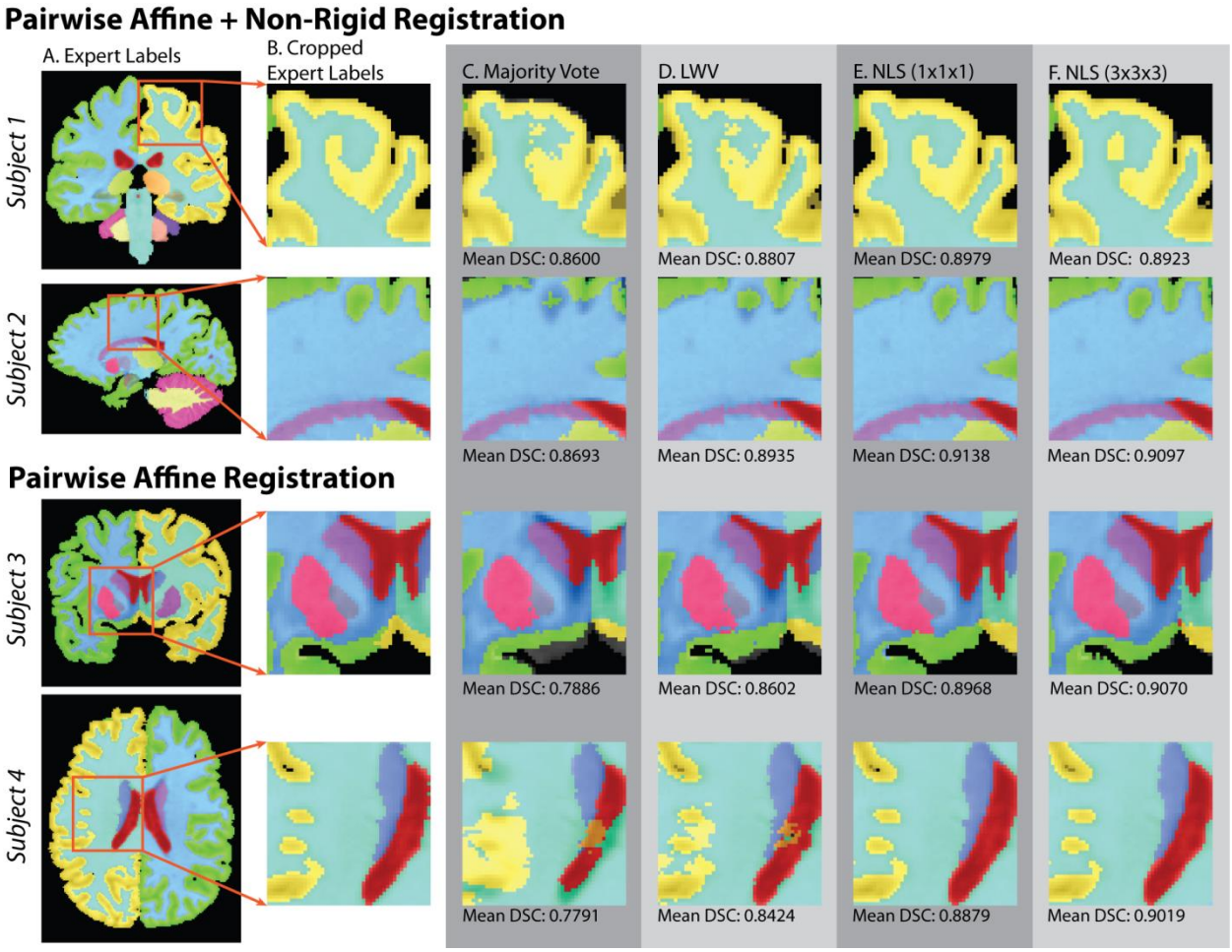


Figure IV.7. Qualitative comparison between the various fusion algorithms for whole-brain segmentation using 5 atlases. For both registration procedures, the qualitative results support the quantitative improvement demonstrated by NLS in Figures IV.4-6. The NLS results are qualitatively superior to alternative voting-based procedures in terms of overall shape, size, location and appearance. Note that the mean DSC labels indicate the mean observed DSC for all labels for the corresponding subject (row) and algorithm (column).

20 out of 25 labels over NLS using a single voxel patch neighborhood. For both registration procedures, none of the baseline algorithms were significantly superior to either NLS approach for any label.

The qualitative results (Figure IV.7) support the quantitative improvement exhibited by NLS over previous algorithms (Figures IV.4-6). Figure IV.7 shows four different subjects (two for non-rigid and two for affine) with the associated expert labels and cropped estimates from the considered baseline algorithms using 5 atlases per estimate. Spatial STAPLE is not shown as it was consistently outperformed by LWV for all considered target images. For reference, MV estimates are provided in order to provide important insight into the quality of the registration. For each presented estimate, the mean DSC value on the presented subject is available below the image. Each example demonstrates the type of improvement exhibited by NLS over voting-based algorithms. NLS provides consistent improvement in terms of shape, size and location of the various labels. Additionally, through the process of finding non-local correspondence, NLS results in segmentation estimates that are qualitatively more consistent in terms of the associated intensity profile, and less dependent upon using high-quality non-rigid registration with large numbers of atlases.

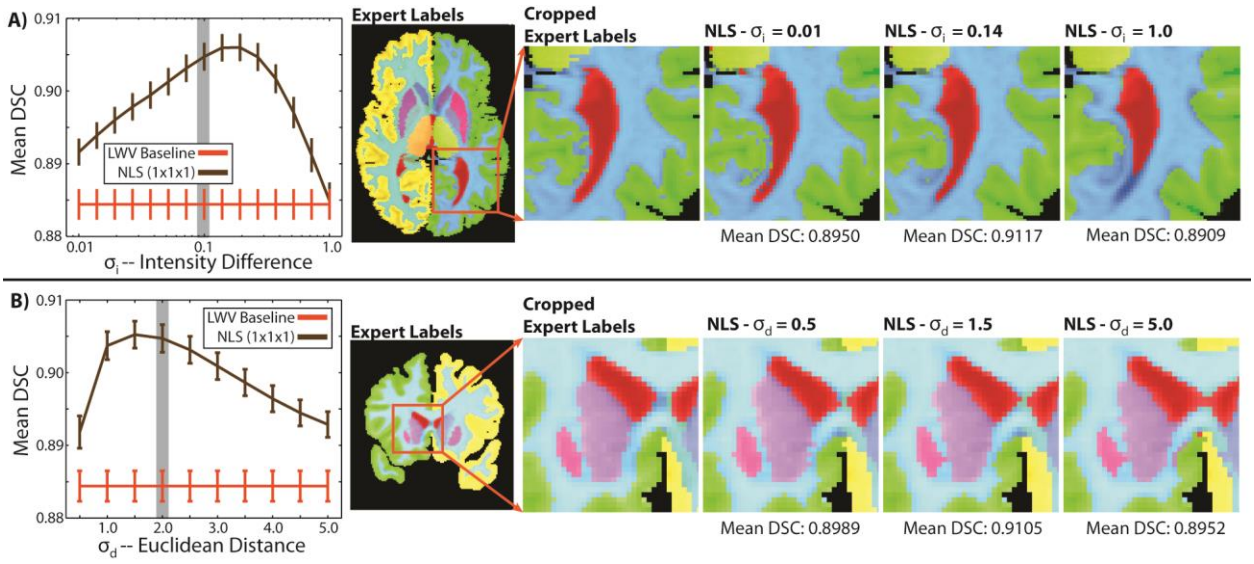


Figure IV.8. Sensitivity to NLS model parameters. The sensitivity of NLS to σ_i (A) and σ_d (B) demonstrate degraded performance for values that are either too small or too large. Regardless, consistent improvement over a locally weighted vote is achieved. Gray outlines indicate the values used in the previously presented experiments. The qualitative results demonstrate the benefits and detriments of optimal and sub-optimal model parameters.

3.7. Sensitivity to Model Parameters

The sensitivity of an algorithm to the model parameters plays a critical role in determining the robustness and applicability of the approach to new problem spaces. The sensitivity of NLS to patch window sizes, quality of registration, and the number of atlases has been presented throughout Figures IV.3-7. Here, we assess the sensitivity of NLS to the two standard deviation parameters, σ_i and σ_d (see Eq. 1). First, σ_i is the standard deviation of the Gaussian intensity difference model and controls how selective the non-local approach is in determining the correspondence between the various voxels. Second, σ_d is the standard deviation of the Gaussian distance model and it weights voxels based upon their distance to the current target voxel of interest. The parameter σ_d can be thought of as a proxy for the size of the search neighborhood (i.e., as the value of σ_d decreases the impact of the extreme elements in the search neighborhood approaches zero). Note, due to this relationship, alternative values for the search neighborhood are not considered. Here, we utilize NLS with a single voxel patch neighborhood on the non-rigidly registered whole-brain data set. Unless the parameter values are being explicitly modified, the previously discussed default parameter values are used.

The results of this sensitivity test (Figure IV.8) demonstrate that NLS is not particularly sensitive to the standard deviation parameters, and continues to exhibit consistent improvement over LWV across a large range of parameter values. Figure IV.8A demonstrates the NLS sensitivity to the σ_i parameter with associated qualitative estimates for various parameter values shown to the right. For values of σ_i that are too small, NLS results in noisy estimates that contain undesired "holes" in the segmentation. On the other hand, large values result in segmentations that are overly smooth and fail to accurately model the underlying intensity profile. While not shown, one important case for this parameter is when $\sigma_i = \infty$ (i.e., ignoring intensity characteristics and only incorporating registration uncertainty via spatial locality). If we set σ_i to ∞ then the algorithm converges to a mean overall accuracy of 0.8746 – an accuracy level statistically indistinguishable from Spatial STAPLE. This provides two important insights (1): it highlights the need of incorporating intensity information into the estimation framework for this particular

application, and (2) it demonstrates that, despite using global performance level parameters, NLS is able to overcome some of the inherent registration uncertainty without directly utilizing the image intensity characteristics. Figure IV.8B shows the sensitivity to the σ_d parameter. In this case, values that are too small cause NLS to use too few voxels to capture the non-local correspondence between the atlases and the target. Values that are too high result in the inclusion of regions of the image that are not anatomically indicative of the label of interest. The gray bars on Figure IV.8 indicate the default values used in the previous experiments.

3.8. Model Optimality

Like STAPLE, NLS is derived in an EM framework in which parameters are iteratively computed in order to estimate the optimal solution for the underlying segmentation. While EM algorithms are guaranteed to converge to a local optimum, convergence to a global optimum is not guaranteed. Thus, it is important to assess the ability of NLS to converge to a reasonable local optimum. Given the true segmentation and a provided non-local correspondence model, it is straightforward to calculate the globally ideal performance level parameters for NLS by replacing the voxelwise label probabilities (i.e., W_{si}) with the true segmentation in Eq. 11

$$\theta^*_{js's} = \frac{\sum_i \left(\sum_{i' \in \mathcal{N}_s(i): D_{i'j}=s'} \alpha_{ji'i}\right) \delta(T_i, s)}{\sum_i \delta(T_i, s)}. \tag{4.12}$$

where $\theta^*_{js's}$ represents the globally ideal performance level parameters, T_i is the true segmentation at voxel i and, $\delta(T_i, s)$ is the Kronecker delta function which is equal 1 if $T_i = s$ and 0 otherwise. For the traditional STAPLE model, the globally ideal performance level parameters can be calculated in a similar manner:

$$\theta^*_{js's} = \frac{\sum_{i: D_{ij}=s'} \delta(T_i, s)}{\sum_i \delta(T_i, s)}. \tag{4.13}$$

Here, for both STAPLE and NLS, we compare the results of the converged algorithm to the results of the algorithm using the globally ideal performance level parameters. We enumerate ideal STAPLE and ideal NLS to indicate the results of the algorithms using the globally ideal performance

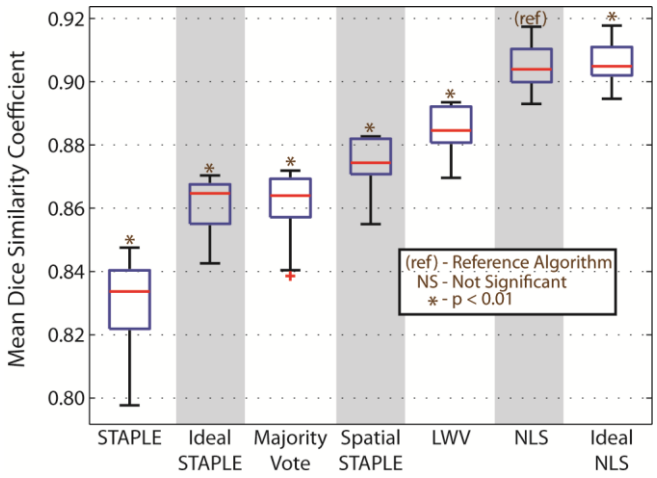


Figure IV.9. Assessment of the model optimality of the NLS approach. The results using ideal STAPLE and ideal NLS represent the estimates using the globally ideal performance level parameters with 5 atlases per estimate. NLS consistently converged to an estimate that is very close to "ideal" NLS (i.e., the global optimum). On the other hand, STAPLE consistently converged to a value significantly less than the global optimum. Additionally, the results of the "Ideal STAPLE" approach are only slightly better than a MV, which indicates the non-optimality of the traditional STAPLE observation model.

level parameters. We assess the results across the 15 whole-brain images using 5 non-rigidly registered atlases per estimate and a single voxel patch neighborhood.

The results of this experiment (Figure IV.9) demonstrate multiple important concepts. The converged NLS estimate is nearly identical to the accuracy of the ideal NLS estimate, which is an indication that, despite using only 5 atlases, NLS is able to converge to an estimate that is very close to the global optimum. To contrast, the converged STAPLE estimate is significantly lower than the ideal STAPLE estimate, which indicates a strong need for using larger numbers of atlases. Additionally, the ideal STAPLE estimate is only slightly better than the MV estimate. Thus, regardless of converging to the global optimum or not, the STAPLE model of rater behavior does not accurately model the observation behavior exhibited in this multi-atlas context. While perhaps surprising, these results are supported by the literature, where, even when large numbers of atlases are used (i.e., the probability of converging to global optimum is increased), STAPLE is, at best, slightly better than a MV in a multi-atlas context [48, 49, 59, 60].

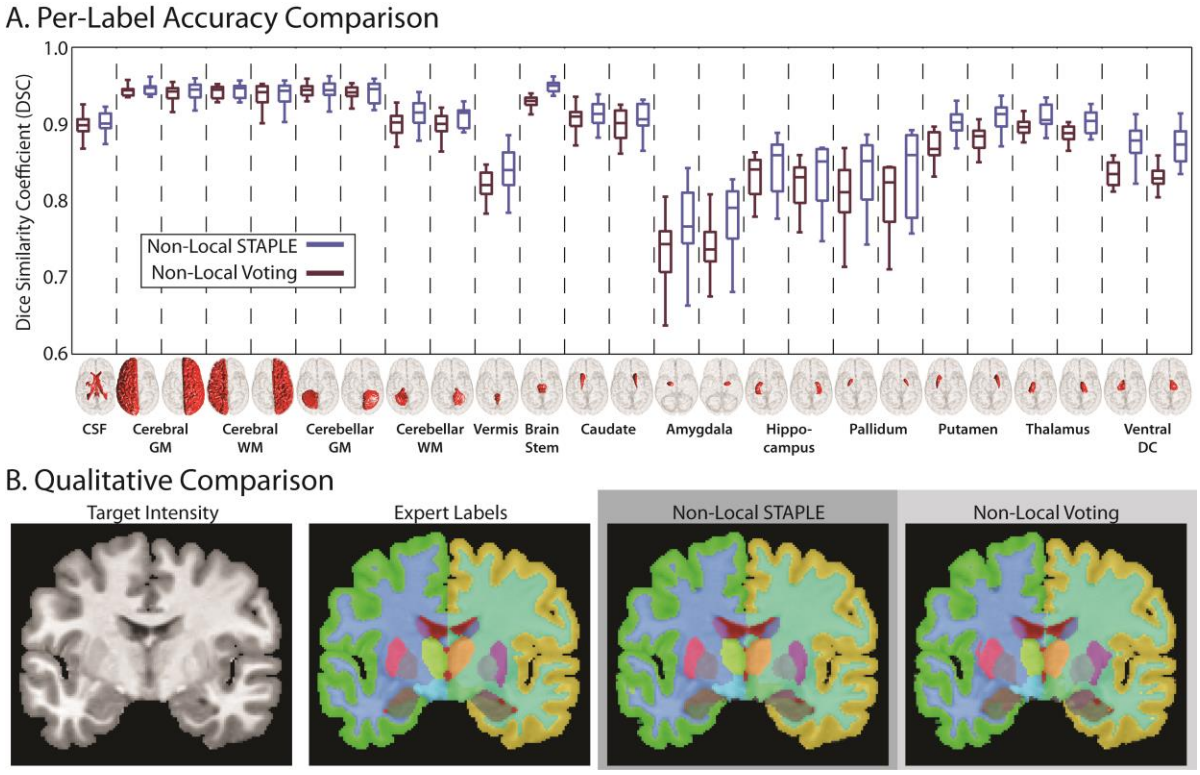


Figure IV.10. Comparison to non-local voting fusion. NLS provided consistent improvement over non-local voting, particularly for the smaller deep brain structures (A). NLS provided significant improvement on 18 of the 25 considered labels. Particularly for the smaller labels, the benefits of the proposed multi-atlas rater model are evident. The qualitative comparison (B) supports the per-label comparison and demonstrates the type of improvement achieved by NLS.

3.9. Comparison to Non-Local Voting

Heretofore, we have limited our comparisons to the algorithms that represent the state-of-the art label fusion algorithms (i.e., the algorithms that are most commonly utilized in the label fusion literature). However, like NLS, recent techniques have been proposed that integrate a non-local correspondence model into a voting-based fusion approach [56, 148]. In order to more fully characterize the performance of NLS to premier segmentation approaches, we compare the results of NLS to a straightforward non-local voting-based procedure [56] for the affine registration whole-brain segmentation problem using 5 atlases per target. For fairness of comparison, identical values were used for NLS and the non-local voting-based approach where applicable (i.e., search neighborhood set to $11 \times 11 \times 11$ voxels, patch neighborhood set to $3 \times 3 \times 3$ voxels, and σ_i set to 0.1).

The results of this comparison (Figure IV.10) indicate that NLS provides significant improvement over non-local voting approaches, particularly when estimating small and more complex deep brain structures. First, a per-label comparison (Figure IV.10A) demonstrates that NLS provides significant improvement ($p < 0.05$, paired t-test) over the non-local voting approach on 18 out of the 25 considered labels. For the larger labels that are more easily distinguishable from the surrounding structures (e.g., CSF, cerebral/cerebellar white and gray matter), NLS and the non-local voting approaches are statistically indistinguishable. However, for the smaller, more complex deep-brain structures (e.g., hippocampus, thalamus, and putamen) NLS provides consistent and significant improvement. The qualitative results (Figure IV.10B) support the quantitative improvement. Here, a representative example from the two approaches is visually presented and NLS is qualitatively superior to the non-local voting approach.

4. Discussion

Non-Local STAPLE represents the first statistical fusion algorithm that seamlessly incorporates intensity into the estimation process and creates a cohesive theoretical model specifically targeting registered atlas observation behavior. Additionally, NLS largely overcomes several of the current obstacles that plague multi-atlas segmentation including the need for high-quality non-rigid registration and large numbers of atlases. These goals are accomplished through the reformulation of the STAPLE algorithm from a non-local means perspective and the integration of the concept of non-local correspondence into the estimation process. Intriguingly, despite this reformulation, the interpretation of the NLS rater model remains straightforward. In words, using a model of non-local correspondence, NLS provides a weighted sum over the non-local search neighborhood to determine what labels *would have been observed* given perfect correspondence between the target and the atlases. Herein, we demonstrated superior performance over state-of-the-art fusion algorithms on two empirical datasets. For thyroid segmentation (Figure IV.3), significant improvement was shown in terms of the DSC, Hausdorff distance, and mean surface distance. For whole-brain segmentation, significant improvement was demonstrated in

terms of overall accuracy (Figure IV.4), per-label accuracy (Figures IV.5 and 6) and qualitative assessment (Figure IV.7).

The sensitivity of the NLS approach was demonstrated with respect to the various model and multi-atlas parameters. In terms of the multi-atlas parameters, NLS is significantly less dependent upon the number of atlases (Figure IV.4) and the quality of the registration (Figures IV.4-7). In terms of the NLS model parameters, the size of the patch neighborhood seems to be particularly dependent upon the quality of registration. For both the thyroid (Figure IV.3) and the pairwise affine whole-brain results (Figure IV.6), where the registration is relatively poor compared to the non-rigid whole-brain registration, patch neighborhoods greater than a single voxel provided significant improvement over smaller patch neighborhoods. Additionally, NLS is fairly insensitive to the two standard deviation parameters in the non-local correspondence model, which further demonstrates the stability of the approach (Figure IV.8). Lastly, and importantly, despite using only 5 atlases, NLS consistently converged to an estimate that is very close to the global optimum (Figure IV.9). While not a definitive proof, this is a strong indication of the optimality of the NLS model of multi-atlas observation behavior.

While the primary focus of this paper is to investigate the theoretical advancements provided by NLS when compared to the state-of-the-art fusion algorithms, we also demonstrate significant improvement over a recently proposed non-local voting-based approach (Figure IV.10). The results of this comparison highlight the benefits of the proposed framework. First, the observed performance increase by NLS is a strong indication that the proposed model of multi-atlas observation error accurately captures empirically observed atlas performance. Second, it indicates a need for the inclusion of a cohesive rater model into the estimation framework so that informed judgment can be made about the applicability of a given atlas to the label estimates, particularly when estimating the complex relationships between easily confused structures. Moreover, while NLS and non-local voting-based approaches similarly include non-local correspondence models, there are stark contrasts in the way in which these techniques estimate the underlying segmentation. In NLS, the non-local correspondence model is used to learn which label an atlas would have observed given perfect correspondence. As a result, all atlases have

an equal opportunity to contribute at all considered voxels, and the quality of an atlas observation is captured by the rater performance parameters. To contrast, in non-local voting, atlases can be completely de-weighted from the estimation process if their intensity characteristics are too different from the target intensity characteristics. As a result, non-local voting-based procedures are susceptible to being biased towards particular atlases and labels as they are more dependent upon accurate intensity normalization and highly representative atlas intensity profiles.

Despite the promise of the NLS fusion model, several questions still persist in order to understand the optimality of the algorithm. For example, the effect of using an alternative similarity metric (e.g., normalized correlation coefficient, mutual information) to the assumed Gaussian difference model presented here (Eq. 1) needs to be investigated. Alternative similarity measures may dramatically lessen the potential impact of noise and the need for accurate intensity normalization between the target and the atlases. Additionally, the procedure for determining the optimal parameter values for a given problem remains primarily *ad hoc*. Statistically driven maximum likelihood and maximum *a posteriori* models to estimate the optimal parameter values through (1) the use of the training data, or (2) direct integration into the estimation model, would provide valuable advancements for the applicability of NLS to new problem spaces.

Additionally, other than Spatial STAPLE, notably absent from the list of considered baseline algorithms are some of the more recent advancements to the STAPLE algorithm. There are two primary reasons for not directly comparing to these extensions. First, it is straightforward to illustrate that NLS is a direct extension of the original STAPLE algorithm. NLS can be thought of as a family of algorithms governed by the non-local correspondence model. From this perspective, the original STAPLE algorithm can be seen as simply a special case of the proposed NLS framework. To illustrate, consider a non-local correspondence model where $\alpha_{ji'i} = 1$ if and only if $i = i'$ and, otherwise, $\alpha_{ji'i} = 0$. In this case, the E- and M-steps (Eq. 7 and Eq. 11, respectively) simplify to the original STAPLE algorithm. Second, we propose that NLS is not mutually exclusive to these proposed advancements. For example, (1) incorporations of spatially varying performance level estimates [49, 52, 54, 67], (2) capturing task

difficulty through the augmentation of the E-step with "consensus levels" [50], (3) locally ignoring atlas voxels based upon *a priori* intensity characteristics [53, 67], and (4) models for stabilizing the performance level parameters [11, 55] could all be seamlessly integrated into the NLS framework. In particular, the recent advancements that allow for local spatially-varying performance level parameters within the STAPLE framework (e.g., Spatial STAPLE and Local STAPLE MAP) represent fascinating potential improvements to the NLS framework. Despite the fact that NLS uses local intensity information in order to reformulate the rater performance model, it remains an inherently global approach as, like the original STAPLE algorithm, the performance level parameters describe global atlas performance. A reformulation of this type of approach to allow for both local intensity characteristics *and* local performance level parameters could potentially provide significant benefit in terms of overall accuracy and robustness. Continued investigation into the integration of the proposed STAPLE advancements represents fascinating avenues of continued research into rater performance model optimality.

CHAPTER V

FORMULATING HIERARCHICAL PERFORMANCE

1. Overview

Label fusion algorithms typically treat all of the considered labels equally. As a result, the complex anatomical relationships that are often exhibited in multi-label segmentation problems are neglected. To illustrate, consider a typical whole-brain segmentation problem in which there are often upwards of 100 unique labels that are estimated. Within those structures there are known anatomical and hierarchical relationships which could be leveraged – e.g., one such relationship might be *medial frontal cortex* → *frontal cortex* → *cerebral cortex* → *cerebrum* → *brain* (where "→" could be interpreted as "is part of"). While generalized hierarchical segmentation frameworks have been around for almost two decades (e.g., [150, 151]) and recently considered for an application-specific voting fusion approach [75], a generalized hierarchical fusion framework has not been considered in the statistical fusion context.

We propose a generalized statistical fusion framework using hierarchical models of rater performance. Building on the seminal STAPLE algorithm, we reformulate the rater performance model to utilize hierarchical relationships through a multi-tier performance model (Figure V.1). The proposed model is built on the simple concept that the performance of a rater at the higher levels of the hierarchical model (e.g., brain vs. non-brain or cerebrum vs. cerebellum) is indicative of the rater's performance at the lower levels of the hierarchy (i.e., the individual labels-of-interest). Thus, the performance at the higher levels of the hierarchy should propagate to lower levels of the hierarchy in a theoretically and probabilistically consistent manner.

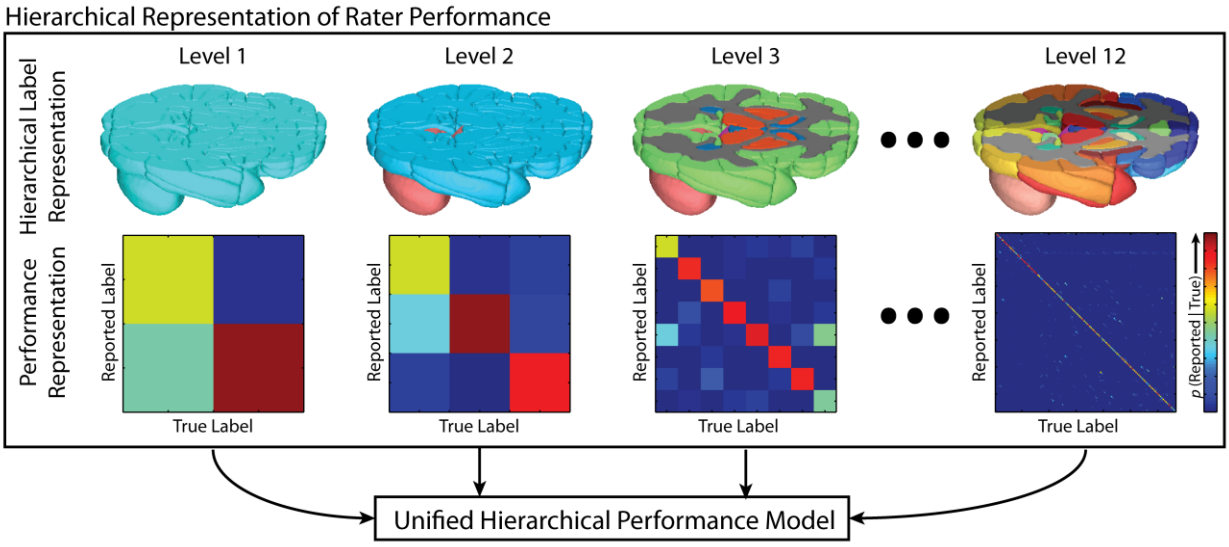


Figure V.1. Hierarchical representation of rater performance. Volumetric renderings of the brain anatomy at the various levels are shown. At each level, the rater performance is quantified using a representative confusion matrix. Each level is then unified through a complete hierarchical performance model.

This chapter is organized in the following manner. First, the theory for the generalized hierarchical statistical fusion framework is derived and the pertinent details for extension to state-of-the-art statistical fusion are provided. Second, we demonstrate superior performance on both simulated and empirical multi-atlas segmentation data – herein, whole-brain and orbital data. Finally, we conclude with a brief discussion on the optimality of the approach and the potential for improvement. The research presented in this manuscript is an extension of a previously published conference paper [79]. Herein, we provide valuable additions to the theoretical derivations and apply the proposed framework to additional simulated and empirical data.

2. Theory

2.1. Problem Definition

Let $\boldsymbol{T} \in \boldsymbol{L}^{N\times 1}$ be the latent representation of the true target segmentation, where $\boldsymbol{L} = \{0, \ldots, L-1\}$ is the set of possible labels that can be assigned to a given voxel. Consider a collection of R raters (or registered atlases) with associated label decisions, $\boldsymbol{D} \in \boldsymbol{L}^{N\times R}$. The goal of any statistical fusion algorithm

is to estimate the latent segmentation, $\boldsymbol{T}$, using the observed labels, $\boldsymbol{D}$, and the provided generative model of rater performance.

2.2. Hierarchical Performance Model

Consider a pre-defined hierarchical model with *M* levels. At each level of the hierarchy, let $\boldsymbol{\mathcal{S}}_m \in \boldsymbol{\mathcal{S}} = \{\boldsymbol{\mathcal{S}}_0, \dots, \boldsymbol{\mathcal{S}}_{M-1}\}$ be a mapping vector that maps a label in the original collection of labels, $s \in \boldsymbol{L}$, to the corresponding label at the m^{th} level of the hierarchy, $\mathcal{S}_{ms} \in \boldsymbol{L}^m$, where $\boldsymbol{L}^m = \{0, \dots, L^m - 1\}$ is the collection labels at the m^{th} level of the hierarchy. Additionally, let the performance of the raters at hierarchical level m be parameterized by $\boldsymbol{\theta}^m \in \mathbb{R}^{R \times L^m \times L^m}$ (i.e., $L^m \times L^m$ *confusion matrix* for each rater). Specifically, $\theta^m_{j\mathcal{S}_{ms'}\boldsymbol{\mathcal{S}}_{ms}}$ is the probability that rater *j* observes label s' given that the true label is *s* at the m^{th} level of the hierarchy. Additionally, let $\boldsymbol{\beta} \in \mathbb{R}^{R \times L}$ be a collection of exponential normalization values that ensure that the generative model is properly normalized. Thus, the generative model is described by

$$f\left(D_{ij} = s' \middle| T_i = s, \boldsymbol{\mathcal{S}}, \{\boldsymbol{\theta}^0, \dots, \boldsymbol{\theta}^{M-1}\}, \boldsymbol{\beta}\right) \tag{5.1}$$

which can be directly interpreted as the probability that rater *j* observes label s' given the true label, hierarchical model, and the corresponding model parameters. To directly estimate this distribution we propose a formulation in which the complete model of hierarchical performance (Eq. 1) is unified through a constrained geometric mean across the multi-tier estimate of rater performance.

$$\begin{aligned} f\left(D_{ij} = s' \middle| T_i = s, \boldsymbol{\mathcal{S}}, \{\boldsymbol{\theta}^0, \dots, \boldsymbol{\theta}^{M-1}\}\right) &= \left(\prod_{m=0}^{M-1} f\left(D_{ij} = \mathcal{S}_{ms'} \middle| T_i = \mathcal{S}_{ms}, \boldsymbol{\theta}^m\right)\right)^{\beta_{js}} \\ &= \left(\prod_{m=0}^{M-1} \theta^m_{j\mathcal{S}_{ms'}\mathcal{S}_{ms}}\right)^{\beta_{js}} \end{aligned} \tag{5.2}$$

where, β_{js} is an exponent that maintains the following constraint:

$$\sum_{s'} \left(\prod_{m} \theta^m_{j\mathcal{S}_{ms'}\mathcal{S}_{ms}}\right)^{\beta_{js}} = 1 \tag{5.3}$$

In other words, β_{js} ensures that the model in Eq. 1 is a valid discrete probability mass function. Note, given the constraints on each individual $\boldsymbol{\theta}^m$ (i.e., a valid confusion matrix) a unique value for β_{js} is

guaranteed to exist and can easily be found using a standard searching algorithm (e.g., binary search, gradient descent). Given the model in Eq. 2 and constraint in Eq. 3, it is now possible to utilize the provided hierarchical model within the statistical fusion EM framework. See Figure V.1 for a graphical representation of the newly proposed generative model of hierarchical performance.

2.3. E-Step: Estimation of the Voxelwise Label Probabilities

Let $\boldsymbol{W} \in \mathbb{R}^{L\times N}$, where $W_{si}^{(k)}$ represents the probability that the true label associated with voxel i is label s at iteration k of the algorithm given the provided information and model parameters

$$W_{si}^{(k)} \equiv f\left(T_i = s \middle| \boldsymbol{D}, \boldsymbol{\mathcal{S}}, \{\boldsymbol{\theta}^0, \dots, \boldsymbol{\theta}^{M-1}\}^{(k)}, \boldsymbol{\beta}^{(k)}\right). \tag{5.4}$$

Using a Bayesian expansion and the assumed conditional independence between the registered atlas observations, Eq. 4 can be re-written as

$$W_{si}^{(k)} = \frac{f(T_i = s)\prod_j f\left(D_{ij} = s' \middle| T_i = s, \boldsymbol{\mathcal{S}}, \{\boldsymbol{\theta}^0, \dots, \boldsymbol{\theta}^{M-1}\}^{(k)}, \boldsymbol{\beta}^{(k)}\right)}{\sum_n f(T_i = n)\prod_j f\left(D_{ij} = s' \middle| T_i = n, \boldsymbol{\mathcal{S}}, \{\boldsymbol{\theta}^0, \dots, \boldsymbol{\theta}^{M-1}\}^{(k)}, \boldsymbol{\beta}^{(k)}\right)} \tag{5.5}$$

where $f(T_i = s)$ is a voxelwise *a priori* distribution of the underlying segmentation. Note that the denominator of Eq. 5 is simply the solution for the partition function that enables $\boldsymbol{W}$ to be a valid probability mass function (i.e., $\sum_s W_{si} = 1$). Using the simplified generative model in Eq. 2, the final form for the E-step of the EM algorithm can be written as

$$W_{si}^{(k)} = \frac{f(T_i = s)\prod_j \left(\prod_m \theta^{m,(k)}_{j\mathcal{S}_{ms'}\mathcal{S}_{ms}}\right)^{\beta_{js}^{(k)}}}{\sum_{s''} f(T_i = s'')\prod_j \left(\prod_m \theta^{m,(k)}_{j\mathcal{S}_{ms'}\mathcal{S}_{ms''}}\right)^{\beta_{js''}^{(k)}}} \tag{5.6}$$

2.4. M-Step: Estimation of the Hierarchical Performance Level Parameters

The estimate of the performance level parameters (M-step) is obtained by finding the parameters that maximize the expected value of the conditional log likelihood function (i.e., using the result in Eq. 6). Unlike the traditional STAPLE approach, however, the parameters for each level of the hierarchy are maximized independently.

$$
\begin{aligned}
\boldsymbol{\theta}_j^{m,(k+1)} &= \arg\max_{\boldsymbol{\theta}_j^m} \sum_i E\left[\ln f\left(D_{ij}=s' \middle| T_i=s, \boldsymbol{\mathcal{S}}, \{\boldsymbol{\theta}^0,\dots,\boldsymbol{\theta}^{M-1}\}, \boldsymbol{\beta}^{(k)}\right) \middle| \boldsymbol{D}, \boldsymbol{\mathcal{S}}, \{\boldsymbol{\theta}^0,\dots,\boldsymbol{\theta}^{M-1}\}^{(k)}, \boldsymbol{\beta}^{(k}\right. \\
&= \arg\max_{\boldsymbol{\theta}_j^m} \sum_i \sum_s W_{si}^{(k)} \ln f\left(D_{ij}=s' \middle| T_i=s, \boldsymbol{\mathcal{S}}, \{\boldsymbol{\theta}^0,\dots,\boldsymbol{\theta}^{M-1}\}, \boldsymbol{\beta}^{(k)}\right) \\
&= \arg\max_{\boldsymbol{\theta}_j^m} \sum_i \sum_s W_{si}^{(k)} \ln\left(\prod_m \theta^m_{j\mathcal{S}_{ms'}\mathcal{S}_{ms}}\right)^{\beta_{js}^{(k)}} \\
&= \arg\max_{\boldsymbol{\theta}_j^m} \sum_i \sum_s W_{si}^{(k)} \beta_{js}^{(k)} \sum_m \ln\theta^m_{j\mathcal{S}_{ms'}\mathcal{S}_{ms}}
\end{aligned}
\tag{5.7}
$$

Noting the constraint that each row of the rater performance level parameters must sum to unity to be a valid probability mass function (i.e., $\sum_{s'} \theta^m_{j\mathcal{S}_{ms'}\mathcal{S}_{ms}} = 1$), we can maximize the performance level parameters at each level of the hierarchical model by differentiating with respect to each element and using a Lagrange Multiplier (λ) to formulate the constrained optimization problem. Following this procedure, we obtain

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \theta^m_{j\mathcal{S}_{ms'}\mathcal{S}_{ms}}}\left[\sum_i \sum_s W_{si}^{(k)} \beta_{js}^{(k)} \sum_m \ln\theta^m_{j\mathcal{S}_{ms'}\mathcal{S}_{ms}} + \lambda \sum_{s'} \theta^m_{j\mathcal{S}_{ms'}\mathcal{S}_{ms}}\right] \\
-\lambda &= \frac{\sum_{i:\mathcal{S}_{mD_{ij}}=\mathcal{S}_{ms'}} \sum_{s'':\mathcal{S}_{ms''}=\mathcal{S}_{ms}} \beta_{js''}^{(k)} W_{s''i}^{(k)}}{\theta^{m,(k+1)}_{j\mathcal{S}_{ms'}\mathcal{S}_{ms}}} \\
\theta^{m,(k+1)}_{j\mathcal{S}_{ms'}\mathcal{S}_{ms}} &= \frac{\sum_{i:\mathcal{S}_{mD_{ij}}=\mathcal{S}_{ms'}} \sum_{s'':\mathcal{S}_{ms''}=\mathcal{S}_{ms}} \beta_{js''}^{(k)} W_{s''i}^{(k)}}{-\lambda} \\
\theta^{m,(k+1)}_{j\mathcal{S}_{ms'}\mathcal{S}_{ms}} &= \frac{\sum_{i:\mathcal{S}_{mD_{ij}}=\mathcal{S}_{ms'}} \sum_{s'':\mathcal{S}_{ms''}=\mathcal{S}_{ms}} \beta_{js''}^{(k)} W_{s''i}^{(k)}}{\sum_i \sum_{s'':\mathcal{S}_{ms''}=\mathcal{S}_{ms}} \beta_{js''}^{(k)} W_{s''i}^{(k)}}
\end{aligned}
\tag{5.8}
$$

where $s'':\mathcal{S}_{ms''}=\mathcal{S}_{ms}$ is the collection of all labels that map to the true label of interest, $\mathcal{S}_{ms}$, and $i:\mathcal{S}_{mD_{ij}}=\mathcal{S}_{ms'}$ is the collection of all voxels in which the observed label, D_{ij}, maps to the observed label of interest, $\mathcal{S}_{ms'}$. At this point, it is important to note: (1) the performance model formulation in Eq. 2 allows for each level of the hierarchy to be maximized independently when maximizing the log-likelihood function, and (2) the result in Eq. 8 uses $\beta_{js}^{(k)}$ which can then be updated, $\beta_{js}^{(k)} \rightarrow \beta_{js}^{(k+1)}$ following the constraint:

$$\sum_{s'}\left(\prod_{m}\theta^{m,(k+1)}_{j\mathcal{S}_{ms'}\mathcal{S}_{ms}}\right)^{\beta^{(k+1)}_{js}} = 1 \tag{5.9}$$

2.5. Extension to state-of-the-art Statistical Fusion Approaches

Recently, there have been several advancements to the statistical fusion framework, for instance (1) characterizing spatially varying performance – Spatial STAPLE [52], (2) incorporation of non-local correspondence models – Non-Local STAPLE (NLS) [78], and (3) a combination of the two – Non-Local Spatial STAPLE (NLSS). In the interest of brevity, we only fully derive the hierarchical version of STAPLE in this manuscript. However, we will briefly describe the extension to each of the above advancements to the statistical fusion framework. Note, while this is certainly not an exhaustive collection of advancements to the statistical fusion framework, the point is demonstrating the amenability of the proposed hierarchical reformulation to the new advancements to the STAPLE framework.

2.5.1. Hierarchical Spatial STAPLE

Background: Spatial STAPLE [49, 52] is an extension to the original STAPLE formulation to allow for smooth voxelwise estimates of rater performance. As a result, the confusion matrix describing rater performance, $\boldsymbol{\theta}_j$, becomes a function of the location in the image, $\boldsymbol{\theta}_{ij}$**,** defined over a pre-defined window surrounding the voxel of interest, $\boldsymbol{B}_i$. Where $\boldsymbol{B}_i$ is the set of voxels that are part of the window (or "pooling region") for the current voxel of interest, i.

E-Step:

$$W^{(k)}_{si} = \frac{f(T_i = s)\prod_j\left(\prod_m \theta^{m,(k)}_{ij\mathcal{S}_{ms'}\mathcal{S}_{ms}}\right)^{\beta^{(k)}_{ijs}}}{\sum_{s''} f(T_i = s'')\prod_j\left(\prod_m \theta^{m,(k)}_{ij\mathcal{S}_{ms'}\mathcal{S}_{ms''}}\right)^{\beta^{(k)}_{ijs''}}} \tag{5.10}$$

M-Step:

$$\theta^{m,(k+1)}_{ij\mathcal{S}_{ms'}\mathcal{S}_{ms}} = \frac{\sigma_{is}\theta^{m,(k)}_{j\mathcal{S}_{ms'}\mathcal{S}_{ms}} + \sum_{i'\in\boldsymbol{B}_i:\mathcal{S}_{mD_{i'j}}=\mathcal{S}_{ms'}}\sum_{s'':\mathcal{S}_{ms''}=\mathcal{S}_{ms}}\beta^{(k)}_{i'js''}W^{(k)}_{s''i'}}{\sigma_{is}\sum_s\theta^{m,(k)}_{j\mathcal{S}_{ms'}\mathcal{S}_{ms}} + \sum_{i'\in\boldsymbol{B}_i}\sum_{s'':\mathcal{S}_{ms''}=\mathcal{S}_{ms}}\beta^{(k)}_{i'js''}W^{(k)}_{s''i'}} \tag{5.11}$$

$$\sum_{s'}\left(\prod_{m}\theta_{ijs_{ms'}s_{ms}}^{m,(k+1)}\right)^{\beta_{ijs}^{(k+1)}}=1 \tag{5.12}$$

where σ_{is} is a regularizing scale factor that biases the local performance estimate to be closer to the global estimate of rater performance (Eq. 8). We formulate σ_{is} to be equal to relative amount that the current label of interest, s, is estimated to be the correct answer over the pooling region, $\boldsymbol{B}_i$:

$$\sigma_{is}=|\boldsymbol{B}_i|-\sum_{i\in\boldsymbol{B}_i}W_{si}^{(k)} \tag{5.13}$$

where $|\boldsymbol{B}_i|$ is the number of elements in the pooling region centered at voxel i. Using this formulation, σ_{is} allows consistent estimates of rater performance despite the fact that only certain labels may be estimated in a given local region of the image – see [52] for further details.

Note, while Spatial STAPLE uses a non-parametric biasing function to prevent instabilities in the local performance estimates, alternative techniques could be used. For instance, one could use a maximum *a posteriori* (MAP) approach and assume a prior Beta distribution on the performance parameters – e.g., [54, 55]. It is straightforward to see that the hierarchical formulation using a MAP formulation would remain valid. Regardless, the optimal framework for characterizing spatially varying performance remains an open problem and outside the scope of this manuscript.

2.5.2. Hierarchical Non-Local STAPLE (NLS)

Background: NLS [51, 78] is another alternative to the original STAPLE algorithm in which the rater performance model is reformulated from a non-local means perspective. Briefly, using the intensity image provided for the target image, $\boldsymbol{I}\in\mathbb{R}^N$, and the corresponding registered intensity images from the atlases, $\boldsymbol{A}\in\mathbb{R}^{N\times R}$, the goal is the estimate the likelihood that $A_{i'j}$ is the true corresponding voxel to the target image I_i, where i' is an element in the search neighborhood defined for voxel i – $\mathcal{N}(i)$. Mathematically, we estimate this likelihood as $f\left(A_{i'j}\middle|I_i\right)$

$$f\left(A_{i'j}\middle|I_i\right)\equiv\alpha_{ji'i}=\frac{1}{Z_\alpha}\Delta\left(A_{i'j},I_i\right)\exp\left(-\frac{\mathcal{E}_{ii'}^2}{2\sigma_d^2}\right) \tag{5.14}$$

where $\Delta\left(A_{i'j}, I_i\right)$ is a generic similarity model, $\exp\left(-\frac{\mathcal{E}_{ii'}^2}{2\sigma_d^2}\right)$ is the spatial compatibility model, and Z_α is a partition function that ensures that $\sum_{i' \in \mathcal{N}(i)} \alpha_{ji'i} = 1$. For the similarity model, there are many different techniques that could be used (e.g., Gaussian difference model [57, 59, 78], locally normalized correlation coefficient [118, 132], mutual information [48]). In the spatial compatibility model, $\mathcal{E}_{ii'}$ is the Euclidean distance between voxels i and i' in image space and σ_d is the corresponding standard deviation. For additional information on NLS see [78].

E-Step:

$$W_{si}^{(k)} = \frac{f(T_i = s) \prod_j \sum_{i' \in \mathcal{N}(i)} \alpha_{ji'i} \left(\prod_m \theta^{m,(k)}_{j\mathcal{S}_{ms'}\mathcal{S}_{ms}} \right)^{\beta_{js}^{(k)}}}{\sum_{s''} f(T_i = s'') \prod_j \sum_{i' \in \mathcal{N}(i)} \alpha_{ji'i} \left(\prod_m \theta^{m,(k)}_{j\mathcal{S}_{ms'}\mathcal{S}_{ms''}} \right)^{\beta_{js''}^{(k)}}} \tag{5.15}$$

M-Step:

$$\theta^{m,(k+1)}_{j\mathcal{S}_{ms'}\mathcal{S}_{ms}} = \frac{\sum_{i:\mathcal{S}_{mD_{ij}}=\mathcal{S}_{ms'}} \overline{\alpha_{Ji}} \sum_{s'':\mathcal{S}_{ms''}=\mathcal{S}_{ms}} \beta_{js''}^{(k)} W_{s''i}^{(k)}}{\sum_i \sum_{s'':\mathcal{S}_{ms''}=\mathcal{S}_{ms}} \beta_{js''}^{(k)} W_{s''i}^{(k)}} \tag{5.16}$$

where $\overline{\alpha_{Ji}} = \left(\sum_{i' \in \mathcal{N}_s(i):\mathcal{S}_{mD_{i'j}}=\mathcal{S}_{ms'}} \alpha_{ji'i} \right)$, and the update of $\boldsymbol{\beta}^{(k)} \rightarrow \boldsymbol{\beta}^{(k+1)}$ is the same as Eq. 9.

2.5.3. Hierarchical Non-Local Spatial STAPLE (NLSS)

Background: NLSS is a unified statistical fusion algorithm that combines the formulations of Spatial STAPLE and NLS into a single unified framework that (1) allows for smooth spatially varying estimates of rater performance, and (2) reformulates the local performance estimates from a non-local means perspective.

E-Step:

$$W_{si}^{(k)} = \frac{f(T_i = s) \prod_j \sum_{i' \in \mathcal{N}(i)} \alpha_{ji'i} \left(\prod_m \theta^{m,(k)}_{ij\mathcal{S}_{ms'}\mathcal{S}_{ms}} \right)^{\beta_{ijs}^{(k)}}}{\sum_{s''} f(T_i = s'') \prod_j \sum_{i' \in \mathcal{N}(i)} \alpha_{ji'i} \left(\prod_m \theta^{m,(k)}_{ij\mathcal{S}_{ms'}\mathcal{S}_{ms''}} \right)^{\beta_{ijs''}^{(k)}}} \tag{5.17}$$

M-Step:

$$\theta^{m,(k+1)}_{ij\mathcal{S}_{ms'}\mathcal{S}_{ms}} = \frac{\sigma_{is}\theta^{m,(k)}_{j\mathcal{S}_{ms'}\mathcal{S}_{ms}} + \sum_{i'\in\boldsymbol{B}_i:\mathcal{S}_{mD_{i'j}}=\mathcal{S}_{ms'}} \overline{\alpha_{Ji'}} \sum_{s'':\mathcal{S}_{ms''}=\mathcal{S}_{ms}} \beta^{(k)}_{i'js''} W^{(k)}_{s''i'}}{\sigma_{is}\sum_s \theta^{m,(k)}_{j\mathcal{S}_{ms'}\mathcal{S}_{ms}} + \sum_{i'\in\boldsymbol{B}_i}\sum_{s'':\mathcal{S}_{ms''}=\mathcal{S}_{ms}} \beta^{(k)}_{i'js''} W^{(k)}_{s''i'}} \tag{5.18}$$

where all of the mathematical formulations are the same as described above for Spatial STAPLE and NLS.

2.6. Initialization, Detection of Convergence and Implementation

Given an *a priori* hierarchical model, there are no additional parameters in the proposed approach when compared to the non-hierarchical implementations of the statistical fusion framework. As a result, the hierarchical statistical fusion algorithms can be initialized in exactly the same was their traditional counterparts. Specifically, for all of the statistical fusion approaches, the performance parameters were initialized by setting the on-diagonal elements to 0.95 and randomly setting the off-diagonal elements to fulfill the required constraints. The voxelwise label prior, $f(T_i = s)$, was initialized using the label probabilities from a "weak" log-odds majority vote (i.e., decay coefficient set to 0.5 voxels) [59]. For Spatial STAPLE, NLSS, and their hierarchical implementations, the pooling region, $\boldsymbol{B}_i$, was set with a half window radius of $5mm$ along all of the principal directions. For NLS, NLSS, and their hierarchical formulations, a Gaussian difference metric (using an intensity standard deviation of 0.1) was used with a half-window radius of $2mm$ along all of the principal directions for both the patch neighborhood and search neighborhood and the spatial standard deviation, σ_d, was set to $1.5mm$

Detection of convergence in the hierarchical statistical fusion framework is slightly different than the traditional approach as we utilize all levels of the hierarchy. Thus, convergence is detected when the normalized trace of the raters' performance parameters at each level of the hierarchy falls below some arbitrary threshold (herein, $\epsilon = 10^{-4}$) between consecutive iterations of the EM algorithm.

$$\frac{1}{LRM}\sum_j\sum_m tr\left(\boldsymbol{\theta}^m_j\right) \tag{5.19}$$

Finally, the implementation of all of the considered statistical fusion algorithms presented in this paper are publicly available as part of the Java Image Science Toolkit (JIST) -- [152], http://www.nitrc.org/projects/jist.

3. Methods and Results

For all of the presented simulations and experiments, the segmentation accuracy is measured using the Dice similarity coefficient (DSC) [140]. Additionally, any claims of statistical significance refer to the results of a Wilcoxon signed-rank test [153] with a p-value threshold of 0.01.

3.1. Motivating Simulation

Before assessing the empirical performance, we present a motivating simulation to demonstrate the manner in which hierarchical models can be integrated into the statistical fusion framework (Figure V.2).

3.1.1. Experimental Design

A single 2D slice model (300×300 voxels, with 7 unique labels) was constructed to loosely approximate the types of relationships that are exhibited in the brain. Given the provided truth model, a collection of 15 labeled observations were constructed by randomly applying boundary errors of varying strength (see Figure V.2A for the best/worst observations). Additional details on the simulation model can be found in [52, 78]. As a baseline, a representative STAPLE result is presented. For incorporating hierarchical models into the statistical fusion framework, a single reference hierarchical structure was established (Figure V.2B). Given, this structure, all unique trees (630 in total) were constructed via label permutation, and the resulting segmentation was estimated using hierarchical STAPLE.

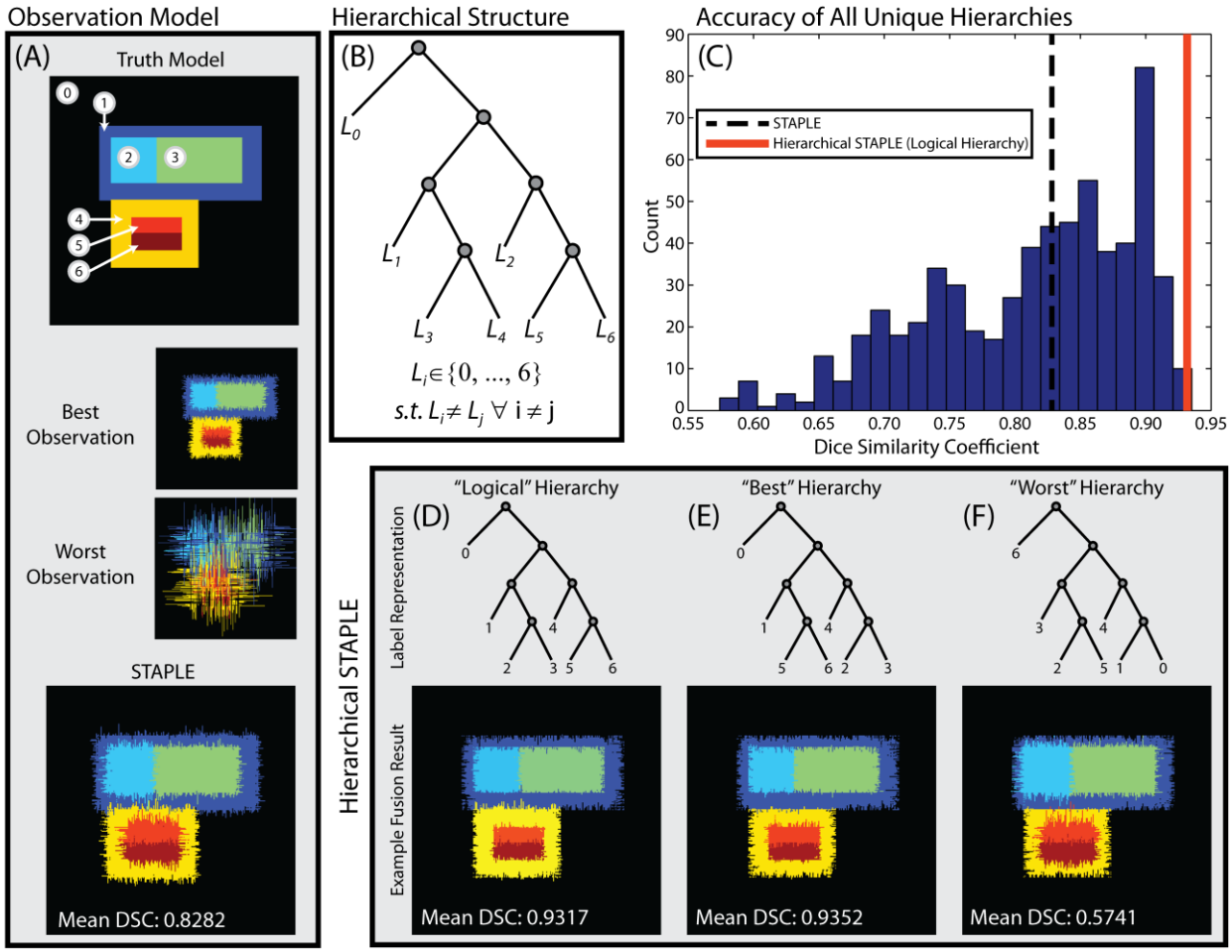


Figure V.2.. Motivating simulation data and results. A simple 2D simulated dataset was constructed with observations using a boundary error model (A). Given a pre-defined hierarchical structure (B), the accuracy of all possible unique hierarchies via label permutation was quantified (C). Representative estimates using the "logical" (D), "best" (E), and "worst" (F) hierarchies are also presented.

3.1.2. Experimental Results

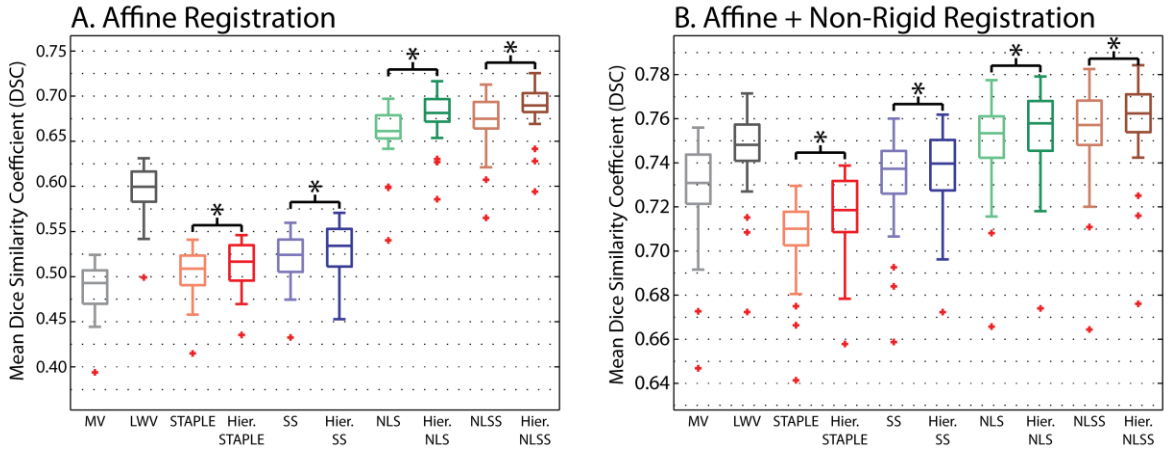
The quantitative results, measured by the mean DSC across 10 Monte Carlo iterations, for each of the considered hierarchical representations can be seen in Figure V.2C. Here, it is evident that the hierarchical label representation plays a substantial role in determining overall segmentation accuracy. For reference, the accuracy of the traditional STAPLE framework and the accuracy using a "logical" hierarchical representation (Figure 2D) are highlighted. The qualitative results (Figure V.2D-2F) support the quantitative assessment of accuracy. Specifically, the accuracy of the "logical" (Figure V.2D), "best" (Figure V.2E), and "worst" (Figure V.2F) hierarchical label representations are presented. While not the

absolute optimal representation in terms of overall mean DSC, the "logical" hierarchical representation: (1) results in a substantial qualitative improvement over the traditional STAPLE estimate and (2) results in a quantitatively superior segmentation estimate than more than 99.5% of the considered hierarchical representations. The "best" hierarchical representation is extremely similar to and results in a very minor improvement over the "logical" representation. Meanwhile, the "worst" representation completely ignores the underlying relationships exhibited in the truth model, and, not surprisingly, results in a very poor estimate of the final segmentation.

3.2. Whole Brain Multi-Atlas Segmentation

3.2.1. Data

For the empirical whole-brain experiments, a collection of 45 MPRAGE images from unique subjects are considered as part of the Open Access Series of Imaging Studies (OASIS, http://www.oasis-brains.org) [145] with subjects ranging in age from 18 to 90. All images had a resolution of $1 \times 1 \times$



Note: * indicates statistical difference using a Wilcoxon signed-rank test ($p < 0.01$).

Figure V.3. Mean accuracy of the various benchmarks and their corresponding hierarchical implementations for both the affine and the non-rigid registration frameworks. The accuracy of a majority vote (MV) and locally-weighted vote (LWV) are presented to provide a reference baseline. The hierarchical implementations for STAPLE, Spatial STAPLE (SS), Non-Local STAPLE (NLS), and Non-Local Spatial STAPLE (NLSS) provide consistent and statistically significant improvement over their non-hierarchical counterparts.

$1mm^3$. All images were labeled using the BrainCOLOR protocol (http://www.braincolor.org/) [154] and provided by Neuromorphometrics, Inc. (Somerville, MA, www.neuromorphometrics.com). Each labeled image contained exactly 133 unique labels (including background). For the purposes of evaluation, 15 of these images were randomly selected as training data, and the remaining 30 were selected as testing data.

3.2.2. Experimental Design

We consider two separate registration frameworks. First we consider an affine-only pairwise registration framework [101] (using "reg_aladin" as part of the "NiftyReg" package – http://sourceforge.net/projects/niftyreg/). Additionally, we consider a pairwise non-rigid registration framework in which the provided affine registrations are augmented with a non-rigid registration [155] (using the Advanced Normalization Tools (ANTs) package – http://stnava.github.io/ANTs/). For both registration frameworks, all 15 training atlases were independently registered to all 30 of the testing atlases – resulting in 450 registrations.

To evaluate fusion performance, we consider several label fusion algorithms. First, in order to provide a benchmark of algorithmic performance, we consider a majority vote [26] and a locally weighted vote (as described in [59]). Additionally, we consider STAPLE [8], Spatial STAPLE [52], NLS [78], and NLSS as well as the hierarchical versions of each, referred to as Hierarchical STAPLE, Hierarchical Spatial STAPLE, Hierarchical NLS, and Hierarchical NLSS, respectively. For the hierarchical algorithms, we constructed a 12-level hierarchical model (manually constructed by an experienced neuroimaging analyst).

3.2.3. Experimental Results

To summarize the improvements exhibited through the use hierarchical performance estimation, the results of the whole-brain multi-atlas segmentation experiment are presented in Figures V.3-V.7. First, to quantitatively summarize the overall improvement for both registration frameworks, the mean DSC (across the 132 non-background labels) is presented in Figure V.3. It is evident that regardless of the registration framework or the specific statistical fusion framework used, the hierarchical reformulation of

the performance parameters provides significant improvement in overall accuracy. For the affine registration framework, the hierarchical implementations provided a mean improvement across the testing data of 0.0070 , 0.0118 , 0.0199 , and 0.0152 for STAPLE, Spatial STAPLE, NLS, and NLSS, respectively. All improvements were statistically significant. For the non-rigid registration framework, the hierarchical implementations provided a mean improvement across the testing data of 0.0104, 0.0049, 0.0048 , and 0.0048 for STAPLE, Spatial STAPLE, NLS, and NLSS, respectively. Again, all improvements were statistically significant. Given the overall improvement in registration quality, the drop in improvement exhibited by the hierarchical implementations for the non-rigid registration framework is expected. Note, for both registration frameworks, the relatively poor performance by STAPLE and Spatial STAPLE are not surprising considering the fact that they do not utilize the atlas-target intensity differences when estimating the final segmentation. Additionally, it is important to note

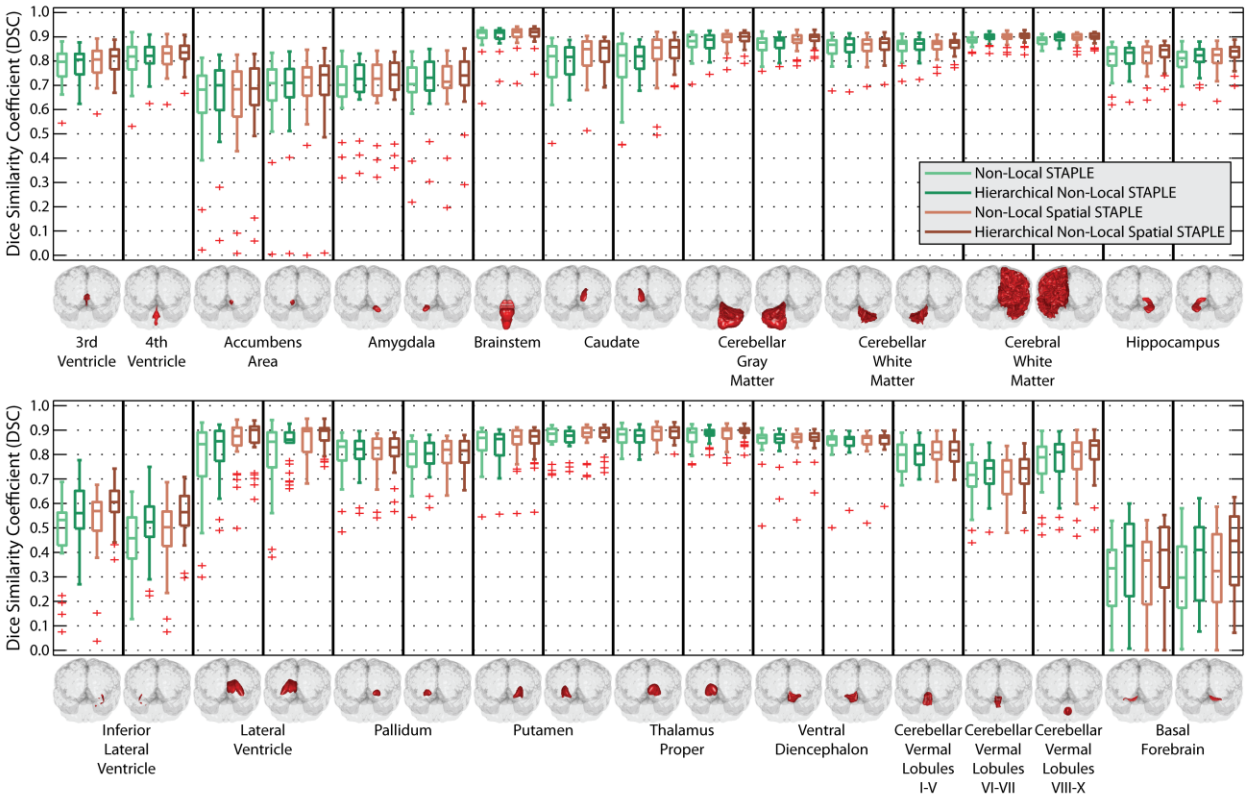


Figure V.4. Per-label accuracy for non-cortical labels for hierarchical implementations of NLS and NLSS using the affine registration framework. The hierarchical reformulations provide substantial and significant improvement for many of the considered labels.

that STAPLE and Hierarchical STAPLE are both out performed by majority vote for the non-rigid registration framework. This highlights the limitations of using a single global performance metric when estimating the final segmentation.

In addition to the overall results, the per-label accuracy for the non-cortical labels using the affine and non-rigid registration frameworks is presented in Figures V.4 and V.5, respectively. Here, for both registration frameworks, only the results using the NLS and NLSS and their hierarchical implementations are presented to avoid obfuscating the improvement provided by their corresponding hierarchical implementations. For the affine registration (Figure V.4), Hierarchical NLS resulted in statistically significant improvement over NLS for 22 of the considered 34 non-cortical labels. Similarly, Hierarchical NLSS resulted in statistically significant improvement for 26 of the 34 non-cortical labels. NLS and NLSS were not significantly superior to their corresponding hierarchical implementations for any of the

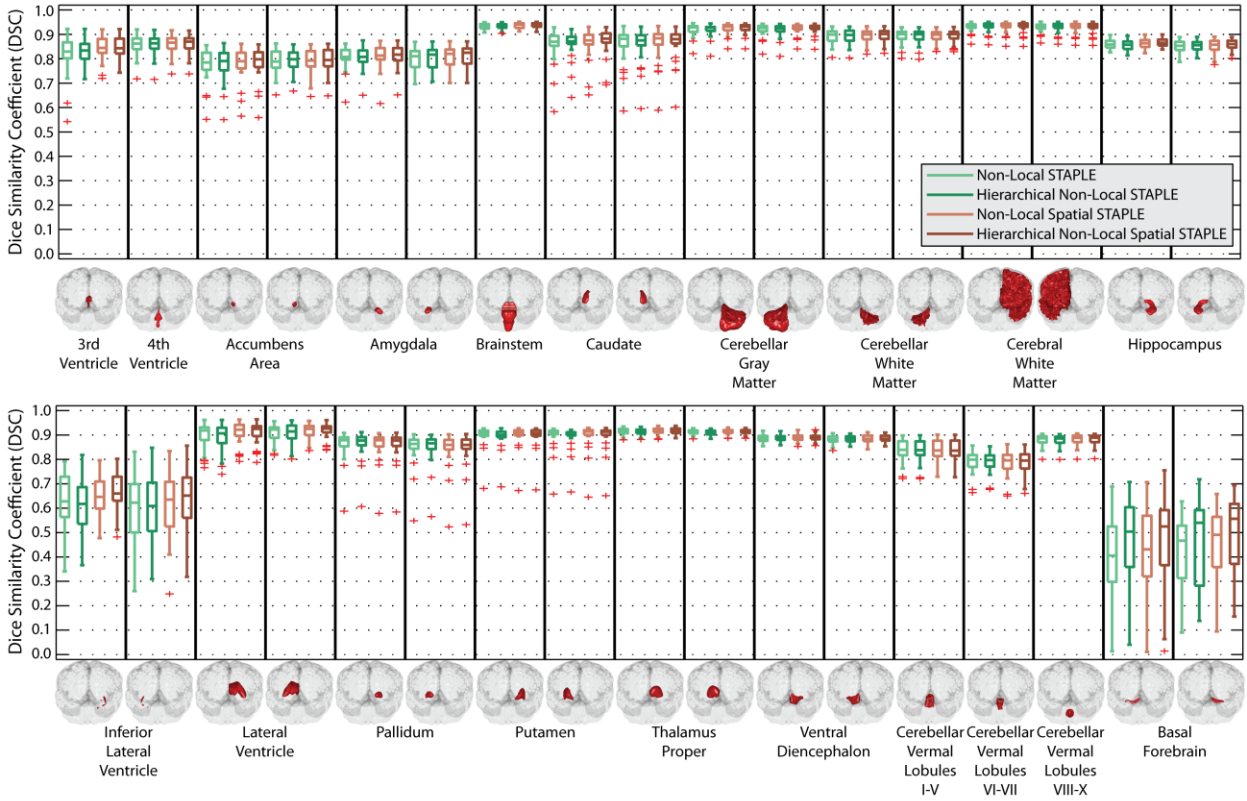


Figure V.5. Per-label accuracy for non-cortical labels for hierarchical implementations of NLS and NLSS using the non-rigid registration framework. As with the affine-only registration framework (Figure V.4), the hierarchical implementations provide substantial and significant improvement for many of the considered labels.

considered labels. For the non-rigid registration (Figure V.5), Hierarchical NLS resulted in statistically significant improvement over NLS for 12 of the 34 considered labels, while Hierarchical NLSS resulted in statistically significant improvement for 19 of the 34 non-cortical labels. As with the affine registration, NLS and NLSS were not significantly superior to their corresponding hierarchical implementations for any of the considered labels.

The quantitative improvement (in terms of the DSC) in the cerebral cortex is summarized in Figure V.6. As with before, only the results using the NLS and NLSS and their hierarchical implementations are presented. Here, it is evident that the hierarchical implementation of each algorithm provides substantial improvement in cortical segmentation accuracy, particularly for the affine-only

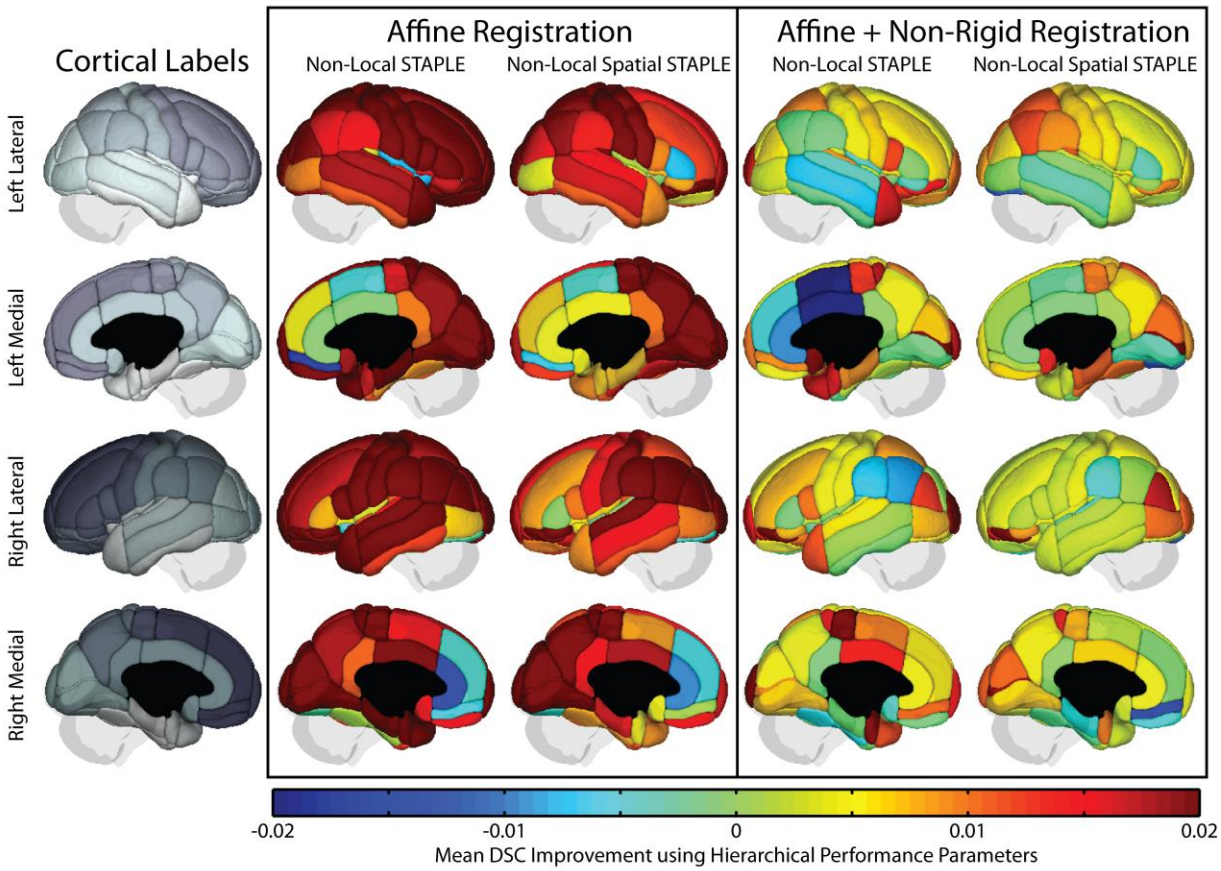


Figure V.6. Mean per-label accuracy improvement for cortical labels using the hierarchical implementations of NLS and NLSS for the both of the considered registration frameworks. Particularly for the affine registration framework, the hierarchical reformulations provide substantial improvement in mean DSC accuracy for many of the cortical labels.

registration framework. To summarize, for the affine registration, Hierarchical NLS resulted in statistically significant improvement over NLS for 57 of the 98 considered cortical labels and was statistically outperformed by NLS on 2 of the 98 cortical labels. Similarly, Hierarchical NLSS resulted in statistically significant improvement for 52 of the 98 cortical labels; however, it was not statistically outperformed by NLSS for any of the considered cortical labels. For the non-rigid registration, Hierarchical NLS resulted in statistically significant improvement over NLS for 28 of the considered 98 cortical labels and was statistically outperformed by NLS on 3 of the 98 cortical labels; while Hierarchical NLSS resulted in statistically significant improvement for 22 of the 98 cortical labels and, again, was not statistically outperformed by NLSS for any of the cortical labels.

The qualitative results (Figure V.7) support the quantitative improvement. Using the affine registration framework, all of the considered statistical fusion algorithms exhibit substantial visual improvement for many of the considered labels. In particular, there appears to be marked improvement in

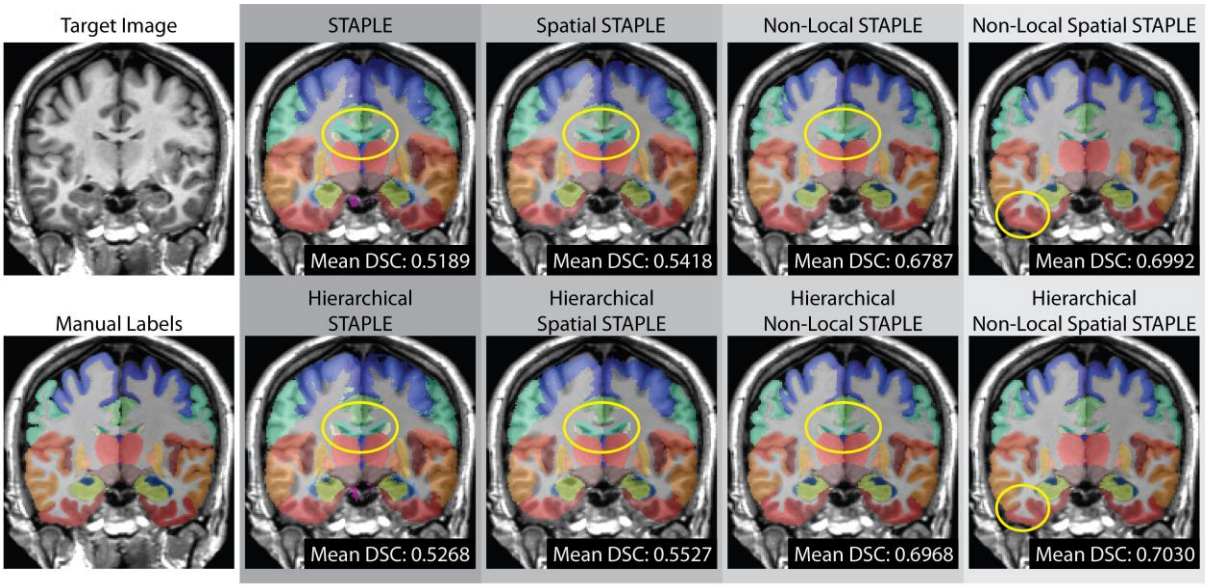


Figure V.7. Qualitative improvement exhibited by several state-of-the-art statistical fusion algorithms with the reformulated hierarchical performance model for the affine registration framework. For each of the considered statistical fusion algorithms we see substantial visual improvement for many of the considered labels. In particular, there appears to be marked improvement in the quality of the lateral ventricle labels and many of the cortical labels. The ellipses highlight regions exhibiting particular qualitative improvement.

the quality of the lateral ventricle labels and many of the cortical labels.

3.3. CT Orbit Multi-Atlas Segmentation

3.3.1. Data

A collection of 31 clinically acquired computed tomography (CT) images of the orbital region were retrieved in anonymous form under IRB supervision. The voxel size of the various images varied wildly, with in-plane resolution of approximately 0.5 mm and slice thickness ranging from 0.4 mm to 5 mm for the various target images. The "ground truth" labels were obtained from an experienced rater and were verified by multiple additional raters. In total, there were 5 considered labels on each dataset: background, left and right optic nerves, and left and right globe/orbital muscles.

3.3.2. Experimental Design

Using a leave-one-out cross-validation (LOOCV) we performed a multi-tier multi-atlas segmentation framework – see [104] for additional details. Briefly, the images were affinely registered [101] and then cropped to form a reasonable region of interest surrounding the orbital area. After cropping, the images were non-rigidly registered [155] and the resulting label conflicts were resolved using label fusion.

Unlike the whole-brain segmentation experiments, the goal of this experiment was two-fold. First, we want to demonstrate the impact of reasonable and logical hierarchical representations of the orbital anatomy on the hierarchical statistical fusion accuracy. To accomplish this, we constructed three logical hierarchical representations (see Figure V.8A). Each of these hierarchical representations could be considered a reasonable representation of the orbital anatomy (e.g., *left optic nerve → optic nerves → non-background* ["Hierarchy 2" in Figure V.8A] or *left optic nerve → left orbit → non-background* ["Hierarchy 3" in Figure V.8A]).

Second, we want to assess the accuracy of the statistical fusion model compared to the "ideal" segmentation estimate (i.e., the segmentation estimate obtained using the "ideal" performance parameters

that are directly calculated using the desired manual segmentation). Obviously, in a typical empirical study, these ideal performance parameters are unknown, and we rely on EM to optimally estimate these parameters. Regardless, given the desired segmentation, obtaining the "ideal" performance parameters is straightforward. For STAPLE, the "ideal" performance parameters are computed as:

$$\theta_{js's}^{(ideal)} = \frac{\sum_{i:D_{ij}=s'} \delta(T_i, s)}{\sum_i \delta(T_i, s)} \tag{5.20}$$

where $\delta(\cdot,\cdot)$ is the Kronecker delta function which is equal to 1 if $T_i = s$ and 0 otherwise. Likewise, for Hierarchical STAPLE the computation of the "ideal" performance parameters is:

$$\theta_{j\mathcal{S}_{ms'}\mathcal{S}_{ms}}^{m,(ideal)} = \frac{\sum_{i:\mathcal{S}_{mD_{ij}}=\mathcal{S}_{ms'}} \sum_{s'':\mathcal{S}_{ms''}=\mathcal{S}_{ms}} \delta(T_i, s'')}{\sum_i \sum_{s'':\mathcal{S}_{ms''}=\mathcal{S}_{ms}} \delta(T_i, s'')} \tag{5.21}$$

where the corresponding exponential normalization factors, $\beta_{js}^{(ideal)}$, are then chosen based upon the following constraint

$$\sum_{s'} \left(\prod_m \theta_{j\mathcal{S}_{ms'}\mathcal{S}_{ms}}^{m,(ideal)} \right)^{\beta_{js}^{(ideal)}} = 1. \tag{5.22}$$

3.3.3. Experimental Results

The results from the LOOCV experiment for multi-atlas segmentation of the orbital region are summarized in Figure V.8. The considered logical hierarchical label representations for this segmentation task are presented in Figure V.8A. The quantitative comparison of STAPLE and the corresponding Hierarchical STAPLE estimates are presented in Figure V.8B. Here, for each of the considered hierarchical representations, Hierarchical STAPLE estimates result in statistically significant improvement over the traditional STAPLE framework. Additionally, the "ideal" Hierarchical STAPLE estimates result in statistically significant improvement over the corresponding "ideal" STAPLE estimate. As a result, it can be directly inferred that *empirically* and *theoretically*, using a reasonable and logical hierarchical representation for Hierarchical STAPLE results in substantial improvement in overall accuracy. Interestingly, "Hierarchy 3" which utilizes the relationships between the left and right orbital

regions, results in best overall performance when estimated using EM and using the ideal parameters. While not definitive, this illustrates the importance of hierarchically grouping labels that are (1) likely to be confused with one another, and (2) indicative of one another's performance.

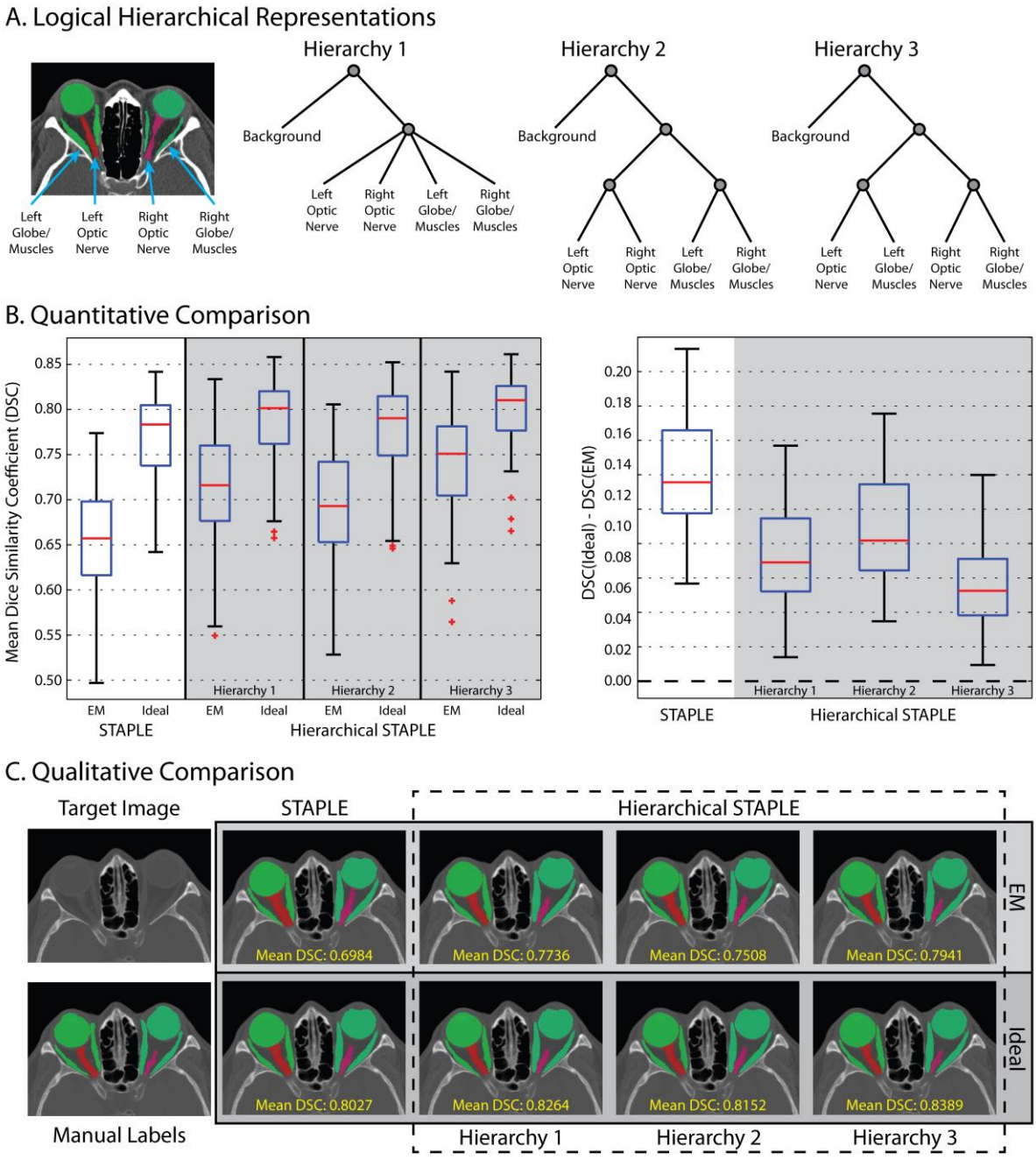


Figure V.8. Empirical evaluation of Hierarchical STAPLE applied to multi-atlas segmentation of orbital anatomy on CT. The considered logical hierarchical representations are shown in (A). The quantitative (B) and qualitative (C) comparisons demonstrate that Hierarchical STAPLE provides significant improvement using both the EM and "ideal" performance parameters.

The “ideal” performance parameters represent an upper bound on potential performance achieved by the statistical fusion framework. In addition to providing statistically significant improvement in overall accuracy over the other considered approaches (Figure V.8B – left), Hierarchical STAPLE using “Hierarchy 3” results in segmentation estimates that are closest the “ideal” performance estimate (Figure V.8B – right). This strongly implies that utilizing a logical representation of the hierarchical relationships exhibited in the data results in an increased likelihood of converging to a local optimum that is closer to the ideal global optimum.

The qualitative results (Figure V.8C) support the quantitative improvement exhibited by Hierarchical STAPLE. Here, it is evident that each of the proposed hierarchical label representations results in: (1) substantial improvement over traditional STAPLE and (2) segmentation estimates that are qualitatively closer to the upper bound provided by the “ideal” segmentation estimate. Again, Hierarchical STAPLE using “Hierarchy 3” results in the largest improvement in mean DSC across the considered labels with an improvement of 0.0957, while “Hierarchy 1” and “Hierarchy 2” result in smaller, yet substantial improvement: 0.0752 and 0.0525, respectively.

4. Discussion

Herein, we propose a novel statistical fusion framework using a reformulated hierarchical performance model. Given an *a priori* model of the hierarchical label relationships for a given segmentation task, the proposed generative model of rater performance provides a straightforward mechanism for quantifying rater performance at each level of the hierarchy. The primary contributions of this manuscript are: (1) we have provided a theoretical advancement to the statistical fusion framework that enables the simultaneous estimation of multiple (hierarchical) confusion matrices for each rater, (2) we have shown that the proposed hierarchical formulation is highly amenable to many of the state-of-the-art advancements that have been made to the statistical fusion framework, and (3) we have demonstrated statistically significant improvement on both simulated and empirical data.

Specifically, through a motivating simulation we have demonstrated the substantial impact that hierarchical label representations have on segmentation accuracy (Figure V.2). For a 133 label whole-brain multi-atlas segmentation task, we have shown substantial and significant accuracy improvement in terms of overall accuracy (Figure V.3), non-cortical segmentation (Figures V.4 and V.5), and cerebral cortex segmentation (Figure V.6). These accuracy improvements are supported by qualitative inspection (Figure V.7). Finally, using a multi-atlas segmentation framework for the orbital region on CT, we evaluated the accuracy of Hierarchical STAPLE using 3 different logical hierarchical representations of the orbital anatomy (Figure V.8). Additionally, using the "ideal" performance parameters as an upper bound, the *empirical* and *theoretical* benefits of the hierarchical performance estimation framework is highlighted.

Despite the promise of the proposed framework, there are several potential advancements that require future exploration. First, all of the presented experiments have relied upon an *a priori* model of the hierarchical relationships within the data. The ability to infer these hierarchical relationships directly from a provided training set would dramatically increase the potential applications for this type of framework, and provide an underlying foundation for estimating the optimal hierarchical formulation for a given application. Second, we have derived this approach from the perspective of hierarchical relationships between labels. However, the same (or very similar) estimation framework could potentially be used to estimate rater performance using multiple labeling protocols. For example, if one had a collection of datasets that were labeled using two separate protocols (either manually or automatically) it may be possible to (1) estimate the relationships between the protocols, and (2) simultaneously estimate rater performance in terms of both protocols. This type of framework is fascinating and certainly warrants further investigation.

In the end, we have presented a powerful theoretical advancement to the statistical fusion context for leveraging the complex inter-structure relationships. While traditional fusion approaches treat all labels equally, the proposed rater model more accurately infers the types of errors that raters (or atlases) make within a hierarchically consistent formulation.

PART 2

APPLICATIONS

The second part of this thesis focuses on translating the previously proposed statistical fusion advancements to clinically and scientifically relevant applications. Herein, we provide three separate applications that build on the theoretical advancements presented in the previous chapters. For the first application (**Chapter VI**), we build a generalized statistical model for anomaly detection (referred to as "out-of-atlas likelihood estimation") that uses fusion-based segmentations to estimate the likelihood of the observed data. For the second application (**Chapter VII**), we build a groupwise multi-atlas segmentation framework to model the shape/appearance of the highly variable internal structure of the spinal cord. Lastly, we derive a machine learning based framework (referred to as "Geodesic Learner Fusion" – **Chapter VIII**) to remove the need for computationally expensive registrations for whole-brain multi-atlas segmentation. These applications highlight the breadth of applications that are enabled by the theoretical foundation governing statistical label fusion and multi-atlas segmentation.

CHAPTER VI

OUT-OF-ATLAS LIKELIHOOD ESTIMATION

1. Introduction

The ability to detect abnormalities and anomalies in medical images plays a critical role in the detection of diseases and pathologies as well as maintaining image quality assurance. A common way to detect abnormalities or anomalies is through the use of a normal template (or atlas) and finding deviations from that template in order to determine the likelihood of an abnormality [156-161]. However, the ability to discover these deviations relies upon the definition of meaningful structure within a target image so that inference can be made about the underlying anatomy. Thus, segmentation plays a critical role in the discovery and quantification of abnormalities and anomalies in medical images.

In multi-atlas segmentation [9, 26], multiple atlases are separately registered to the target and the voxelwise label conflicts between the registered atlases are resolved using label fusion. In general, there are two primary fields of study in label fusion. (1) voting-based strategies which include a majority voting [26, 62-64] and weighted voting strategies [48, 56, 57, 59-61, 72] and (2) statistical fusion strategies based upon Simultaneous Truth And Performance Level Estimation (STAPLE) [8] and the proposed extensions [9, 11, 49-52, 54, 55, 67]. Multi-atlas segmentation has been shown to be highly robust across an extraordinary range of potential applications (e.g., segmentation of the thyroid [51], hippocampus [60], neonatal brain anatomy [162], and the optic nerve [163]).

Nevertheless, there are two primary concerns that limit the generalizability of multi-atlas segmentation. Firstly, we are limited to structures that are represented by the atlases — multi-atlas segmentation cannot be used segment structures that are not present on the available atlases. Secondly, we are limited to structures that are anatomically consistent across potential target subjects. For example,

regardless of whether there are atlases available, a direct multi-atlas segmentation procedure cannot be used to segment malignant gliomas in the human brain as tumor characteristics (e.g., location, size, shape) are widely varying across a given target population. As a result, the potential scope of multi-atlas segmentation applications is limited, particularly in the case of anatomical abnormalities (e.g., the detection of highly-varying pathologies) and quality control (e.g., the detection of imaging and quality-based artifacts). We enumerate this problem as the fact that multi-atlas segmentation is limited to "in-

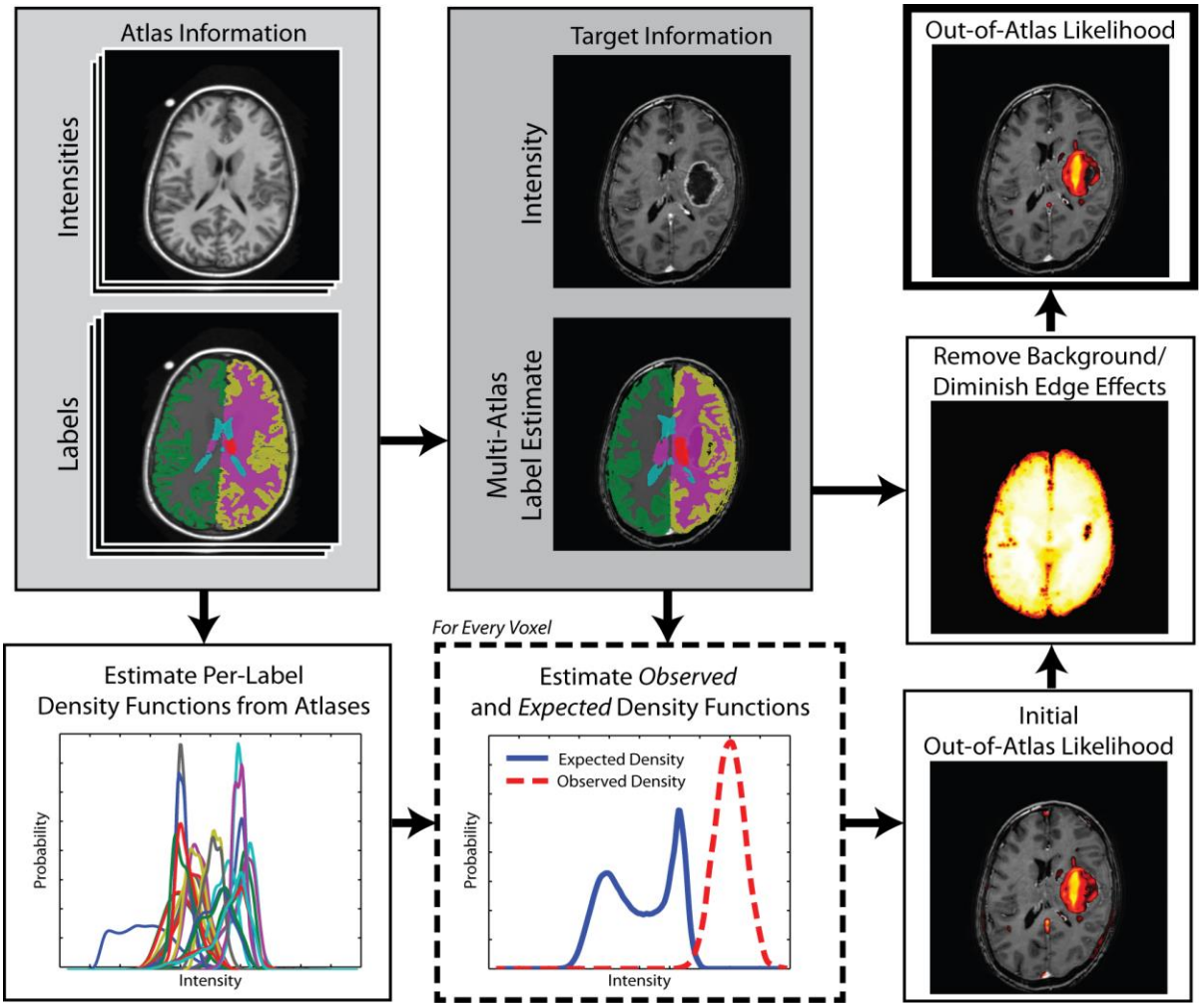


Figure VI.1. Flowchart demonstrating the out-of-atlas likelihood estimation procedure. First the provided atlas information is used to both (1) perform a multi-atlas segmentation estimate of the target image, and (2) estimate the per-label density functions. Next, these per-label density functions and the target information are used to estimate the *observed* and *expected* density functions. These two density functions are then used to construct a voxelwise estimate of the out-of-atlas likelihood. Lastly, the background and edge effects are diminished through a post-processing smoothing step.

atlas" applications (e.g., applications where the atlases are anatomically and structurally indicative of the target image).

Herein, we propose a technique to estimate the out-of-atlas (OOA) likelihood for every voxel in the target image (Figure VI.1). The OOA approach provides an intuitive and fully general abnormality/outlier detection framework that (1) overcomes several of the current limitations with multi-atlas segmentation and (2) has the potential to dramatically increase the scope of potential multi-atlas-based applications.

This chapter is organized as follows. We begin by deriving the theoretical basis and the model parameters for the proposed OOA likelihood estimation framework. Next, using a collection of manually labeled whole-brain datasets, we demonstrate the efficacy of the proposed framework on two distinct applications. First, we demonstrate the ability to detect malignant gliomas in the human brain -- an aggressive class of central nervous system neoplasms. For this application, we both quantitatively and qualitatively assess the accuracy of the proposed algorithm and demonstrate its sensitivity to the various model parameters and initializations. Second, we demonstrate how this OOA likelihood estimation framework can be used within a quality control context for Diffusion Tensor Imaging (DTI) datasets. Using a clinically acquired dataset, we qualitatively demonstrate that we can detect large-scale quality control issues (e.g., aliasing, shading artifacts) within the proposed estimation framework.

2. Theory

In the following presentation of theory we derive the theoretical basis for the OOA likelihood estimation framework and provide a brief overview of the model parameters and initialization procedure.

2.1. Problem Definition

Consider an image of N voxels with unknown target labels $\boldsymbol{T}$, $T_i \in \{0, 1\}$ (i.e. 0: "in-atlas" and 1: "out-of-atlas"). R registered atlases (or "raters" in common fusion terminology) each provide an observed delineation of all N voxels exactly once. The set of labels on these atlases, $\boldsymbol{L}$, represents the set of possible

values that an atlas can assign to all N voxels. Let $\boldsymbol{D}$ be an N × R matrix that indicates the label decisions of the R registered at all N voxels where each element $D_{ij} \in \{0, 1, \dots, L-1\}$. Let A be another N × R matrix that indicates the associated post-registration atlas intensities for all R atlases and N voxels where $A_{ij} \in \mathbb{R}$. Lastly, let $\boldsymbol{I}$: $I_i \in \mathbb{R}$ be the N-vector representing the target intensities, and let $\boldsymbol{\Psi}$: $\Psi_{\mathrm{i}} \in \boldsymbol{L}$ be the N-vector representing the multi-atlas segmentation estimate of the target image.

2.2. Construction of the Expected Intensity Distributions

We define the *expected* intensity distribution as the approximate semi-local intensity distribution that would be observed given the provided atlas label-intensity relationships and the multi-atlas segmentation estimate at each voxel on the target image. This *expected* intensity distribution is approximated by summing the observed label-intensity relationships from the atlases across the multi-atlas segmentation estimate of the target within the semi-local neighborhood around the current voxel of interest. Thus, the first step is to construct the label-intensity relationships that can be inferred from the provided atlas information. In other words, we need to construct $p(\gamma|\Psi_{\mathrm{i}} = l)$ which represents the probability of all possible intensities given that the estimated label is l. We infer this distribution fully from the atlas intensities and labels using a non-parametric Kernel Density Estimation (KDE) approach

$$p(\gamma|\Psi_i = l) = \frac{\sum_j \sum_{i:D_{ij}=l} K\left(\frac{\gamma - A_{ij}}{h}\right)}{h \sum_j \sum_i \delta(D_{ij}, l)} \tag{6.1}$$

where γ is all possible intensities, K is a standard Gaussian kernel, and h is the bandwidth associated with the Gaussian kernel, and δ is the Kronecker delta function. Given Eq. 1, which is an estimation of the complex label-intensity relationships inferred from the atlases, the *expected* intensity distribution within a semi-local neighborhood can then be estimated using the multi-atlas segmentation estimate of the underlying target image

$$p_i^E(\gamma) = \frac{1}{Z_i^E} \sum_{i' \in \mathcal{N}_i} p(\gamma|\Psi_{i'}) \tag{6.2}$$

where $\mathcal{N}_i$ is the semi-local neighborhood surrounding the target voxel i and Z_i^E is the partition function that enforces that $p_i^E(\gamma)$ is a valid probability density function across all potential image intensities. In other words, Z_i^E enforces the constraint that

$$\int_{-\infty}^{+\infty} p_i^E(\gamma)\, d\gamma = 1. \tag{6.3}$$

2.3. Construction of the Observed Intensity Distributions

We define the *observed* intensity distribution at a given target voxel as simply the KDE of the intensities within a semi-local neighborhood surrounding the current voxel of interest on the target image. The *observed* intensity distribution is approximated using a similar approach to Eq. 1

$$p_i^O(\gamma) = \frac{\sum_{i \in \mathcal{N}_i} K\left(\frac{\gamma - I_i}{h}\right)}{h|\mathcal{N}_i|} \tag{6.4}$$

where $\mathcal{N}_i$, K, and h are defined in the same way as Eqs. 1 and 2, and $|\mathcal{N}_i|$ is the cardinality of the set $\mathcal{N}_i$ (i.e., the number of elements in the semi-local neighborhood).

2.4. Estimation of the Voxelwise Out-of-Atlas Likelihood

We define the out-of-atlas likelihood as the voxelwise difference between the *expected* and the *observed* intensity distributions. There are several potential techniques that could be used to capture the difference between these two density functions (e.g., Kullback-Leibler Divergence [164]). Here, we have found that the best way to capture the difference between these distributions is by integrating over the intensities by which $p_i^O(\gamma)$ is greater than $p_i^E(\gamma)$ (i.e., the intensities for which the *observed* probabilities are greater than the *expected* probabilities). Mathematically, this quantity is defined as:

$$p(T_i = 1) = \mathcal{L}_i = \int_{-\infty}^{+\infty} I\left(p_i^O(\gamma) > p_i^E(\gamma)\right)\left[p_i^O(\gamma) - p_i^E(\gamma)\right] d\gamma \tag{6.5}$$

where $\mathcal{L}_i$ represents the OOA likelihood at target voxel i, and $I(\cdot)$ is the indicator function. This formulation of the OOA likelihood has several benefits. First, it is guaranteed that the value of $\mathcal{L}_i \in [0, 1]$ given that $p_i^O(\gamma)$ and $p_i^E(\gamma)$ are properly normalized density functions. Second, this formulation has an

easily understood probabilistic interpretation where $\mathcal{L}_i = 1$ indicates an out-of-atlas likelihood of unity, and $\mathcal{L}_i = 0$ indicates an out-of-atlas likelihood of zero.

2.5. Model Parameter Initialization, and Implementation Details

There are two primary model parameters that need to set in order to use the OOA likelihood estimation framework: (1) the neighborhood structure, $\mathcal{N}_i$, and (2) the bandwidth for the KDE formulation. First, for all presented experiments we used an approximately $11 \times 11 \times 11$mm window centered at the target voxel of interest for all voxels within the neighborhood structure. For cases where this this window size resulted in fractional number of voxels in a given direction, the number of voxels was rounded appropriately. Second, unless otherwise noted, the bandwidth parameter, h, was set to 1.0. Note that this parameter is inherently related to the variance of the observed data, and, thus, a function of the intensity normalization process.

Additionally, one extremely important aspect of this algorithm is the way in which the multi-atlas segmentation estimate is acquired. For all presented experiments, all atlases were registered to the target image using a pairwise registration procedure (i.e., all atlases were independently registered to the target). The intensities between the target and the atlas images were normalized in a two-step process. First, both the target and the registered atlas images are normalized so that the intensities are distributed as a unit Gaussian distribution within the brain region. Second, a second order polynomial is fit to each atlas by finding a least squares solution for the polynomial coefficients that map the mean of each label on the target (via an initial majority vote) to the corresponding labels on the atlases. Lastly, the registered atlases were then fused using Non-Local STAPLE (NLS) [51]. For all presented experiments, NLS was initialized with performance parameters equal to 0.95 along the diagonal and randomly setting the off-diagonal elements to fulfill the required constraints. For all presented results, the voxelwise label prior was initialized using the probabilities from a "weak" log-odds majority vote (i.e., decay coefficient set to 0.5) [59], the search neighborhood, $\mathcal{N}_s(i)$, was initialized to an $11 \times 11 \times 11$ mm window centered at the target voxel of interest, and the patch neighborhood, $\mathcal{N}_p(\cdot)$, was initialized to a $3 \times 3 \times 3$ mm window.

The values of the standard deviation parameters, σ_i and σ_d, were set to 0.1 and 3, respectively. Consensus voxels were ignored during the estimation process. Convergence of NLS was detected when the average change in the trace of the performance level parameters fell below 10^{-4}. For a full derivation of NLS and additional details on NLS initialization, we refer the reader to [51].

Lastly, there are a couple of important implementation details that need to be discussed. First, the OOA likelihood, $\mathcal{L}_i$, was only calculated on voxels for which the multi-atlas segmentation estimate was non-background. Background voxels were ignored because (1) as both of our empirical experiments are for whole-brain analysis, it is assumed that we are only interested in abnormalities that take place within the brain region and (2) it dramatically decreases the runtime of the algorithm. Secondly, a post-processing step that decreased the potential edge effects on the image was performed. Due to the fact that we are ignoring background voxels, it is possible that undesired likelihood estimates could be achieved along the boundaries between background and non-background voxels. To alleviate this problem, we multiplied the final OOA likelihood estimate by an inverse log-odds estimate (decay coefficient set to 1.0 [59]) of the background label (see Figure VI.1 for a visual representation of this process).

3. Methods and Results

We present two starkly different whole-brain empirical experiments in order to assess the efficacy of the proposed OOA likelihood estimation framework. For our first experiment, we both quantitatively and qualitatively assess the ability of our framework to detect malignant gliomas in the human brain based on clinically acquired MRI data. Additionally, we provide insight into the sensitivity of the proposed framework with respect to the KDE bandwidth parameter and the accuracy of the multi-atlas segmentation estimate. For our second experiment, we provide a qualitative example for how this OOA model could be used to provide a quality control framework for acquired DTI images and demonstrate the type of imaging artifacts and quality control metrics that could be performed.

3.1. Multi-Atlas Data

The collection of whole-brain atlases used in the following experiments is a collection of 15 T1-weighted Magnetic Resonance (MR) images of the brain as part of the Open Access Series of Imaging Studies (OASIS) [145] dataset. This data was expertly labeled courtesy of Neuromorphometrics, Inc. (Somerville, MA) and provided under a non-disclosure agreement. A refined dataset (using the OASIS brains and a subtly revised labeling protocol) has recently been made available as part of the MICCAI 2012 workshop on multi-atlas labeling. This data is available at the following URL: https://masi.vuse.vanderbilt.edu/workshop2012/ or directly from Neuromorphometrics. For each atlas, a collection of 26 labels (including background) were considered: ranging from large structures (e.g., cortical gray matter) to smaller deep brain structures. Note that all of the cortical surface labels were combined to form left and right cortical gray matter labels. All images are 1mm isotropic resolution.

3.2. Detection of Malignant Gliomas

Thirty pre-operative gadolinium-enhanced T1-weighted brain MRI scans based on varied (but standard of care) imaging protocols with malignant gliomas were obtained in anonymous form under Institutional Review Board (IRB) approval. On average, the resolution of each of the patient image is $0.45 \times 0.45 \times 3\text{mm}$. The corresponding "ground truth" labels associated with each of the tumor regions were manually drawn using the Medical Image Processing And Visualization software [165].

For each target image, all pairwise affine registrations between the 15 labeled atlases and the target image were performed using FLIRT (FMRIB, Oxford, UK). Note that non-rigid registration was not performed due to the highly variable imaging characteristics of the malignant gliomas on the target subjects. We assess the quantitative accuracy of the proposed approach by analyzing the positive predictive value (PPV), negative predictive value (NPV) and the corresponding Receiver Operating Characteristic (ROC) associated with each target image for varying threshold values of the estimated OOA likelihood. All quantitative results are presented in reference to the corresponding manual labels.

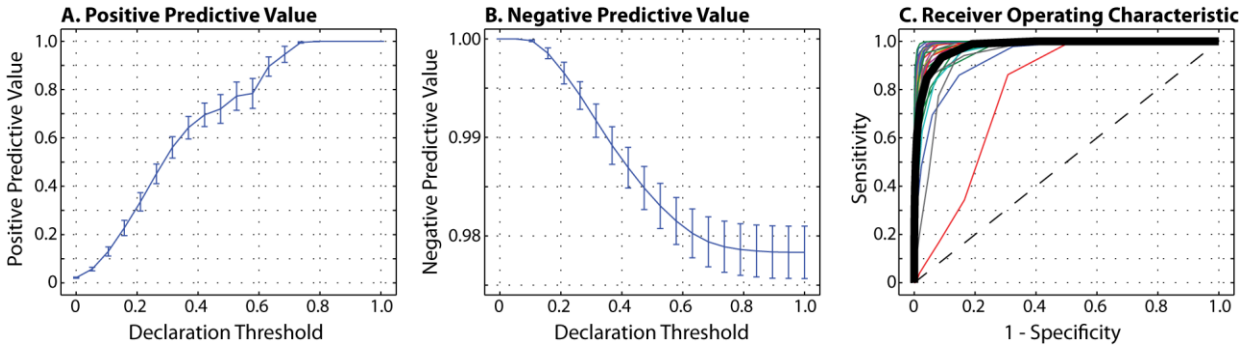


Figure VI.2. Quantitative results for the detection of malignant gliomas across 30 target subjects. The positive and negative predictive values for varying declaration thresholds can be seen in (A) and (B), respectively. The "declaration threshold" indicates the threshold probability for which we declare a voxel to be anomalous (in this case, a cancerous voxel). Finally, the per-subject Receiver Operating Characteristic (ROC) curves can be seen in (C) in the various thin lines, with the mean ROC curve across the subjects represented with the thick black line.

Additionally, we present the sensitivity of the approach to the KDE bandwidth parameter, and various multi-atlas label fusion approaches.

3.3. Glioma Detection Results

The quantitative results (Figure VI.2) demonstrate the ability of the proposed framework to detect large-scale abnormalities in the human brain. The proposed framework can consistently and reliably declare voxels to be cancerous in terms of increasing declaration threshold (Figure VI.2A). For a declaration threshold above approximately 0.7 the resulting PPV was equal to unity (i.e., all voxels declared to be OOA were cancerous voxels). To support these PPV values, the NPV values (Figure VI.2B) show that, despite increasing the threshold, the negative predictive value remains over 0.97. The per-subject ROC curves (Figure VI.2C) confirm that this performance is consistent across the target population with an average area under curve (AUC) value of greater than 0.95. Qualitative results (Figure VI.3) support the quantitative accuracy. While the resulting likelihood estimates are far from perfect (e.g., "holes" in the likelihood estimates), it is evident that the proposed framework is consistently detecting the cancerous regions. The representative example in the fifth column represents the worst-case of the considered subjects, and it is shown that while none of the image has an OOA likelihood of greater than

0.6, the values greater 0.3 are outside of the "core" of the glioma. Note that this example is represented by the outlier case in Figure VI.2C.

The OOA approach is not particularly sensitive to the bandwidth parameter (Figure VI.4), with the optimal setting being approximately 1.0. Note that this value is not particularly surprising as the atlas and target images were normalized to a unit Gaussian distribution as part of the pre-processing steps. The qualitative results in Figure VI.4B-4G demonstrate the effect of the various bandwidth values. For values that are too small (e.g., 0.5) largely normal regions of the anatomy are declared OOA, while, for values

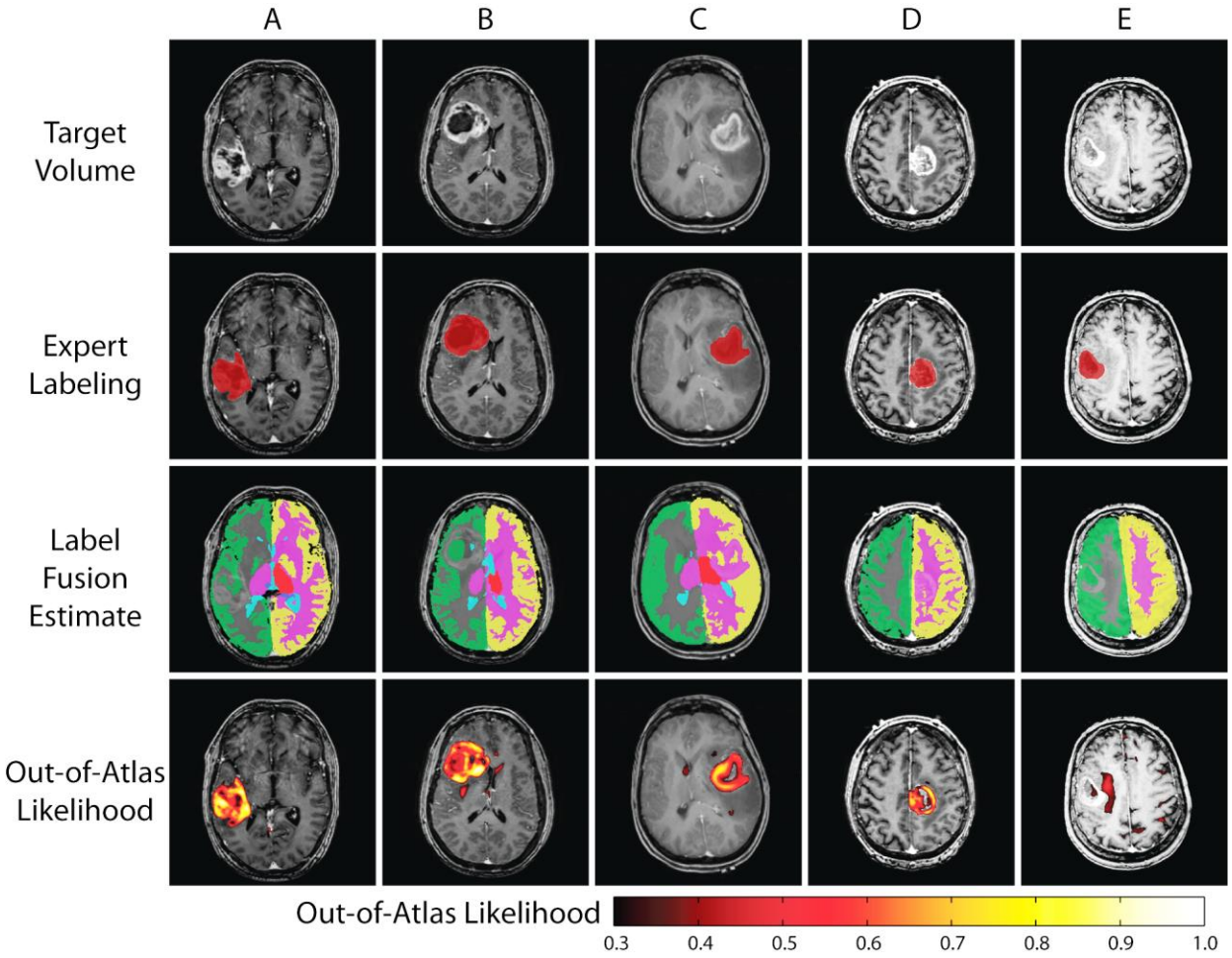


Figure VI.3. Qualitative results for the detection of malignant gliomas. Five representative examples are presented. For each example, the target volume, expert labeling, label fusion estimate, and the out-of-atlas likelihood are presented. The first four examples represent cases where the tumor region is correctly identified. The last example represents the outlier case (seen in red in Figure VI.2C) in which the cancerous region was almost completely missed.

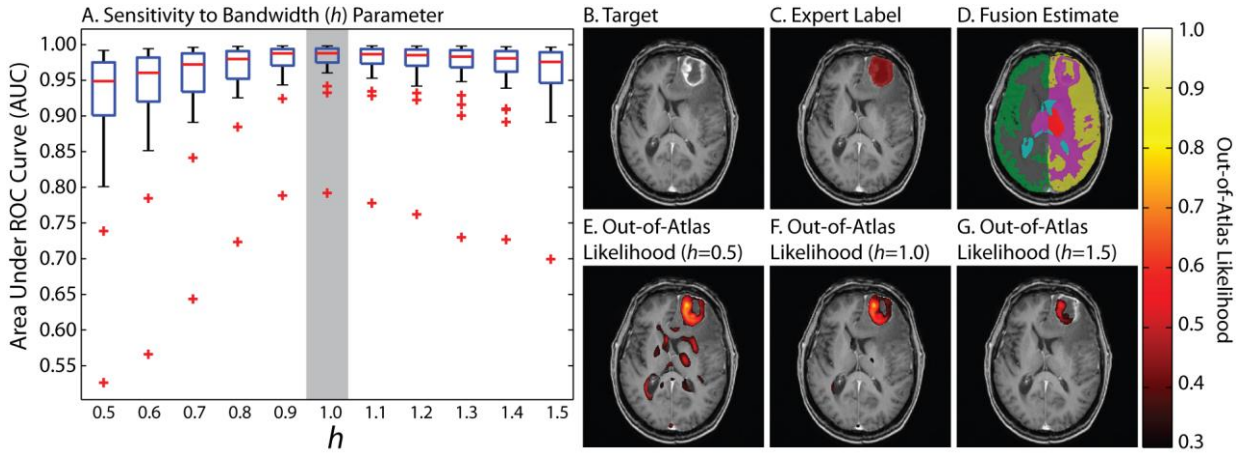


Figure VI.4. Sensitivity of the approach to the bandwidth parameter. The spread of area under curve (AUC) values across the 30 subjects for various bandwidth values is seen in (A). Note that the optimal value is approximately 1.0, which is not surprising given the intensity normalization procedure. In (B)-(G) qualitative results are presented with various out-of-atlas likelihood estimations for varying bandwidth values presented in (E)-(G).

that are too large (e.g., 1.5), the OOA likelihoods are not strict enough and fail to discover a large portion of the malignant glioma. Fusion of multiple atlases consistently outperforms using the best individual atlas (in terms of the mean ROC curve across the target population - Figure VI.5). NLS (which utilizes the intensity information of the atlas-target relationships) consistently results in more accurate labels, and, thus, more accurate OOA likelihood estimates than traditional STAPLE and a majority vote fusion approaches.

3.4. DTI Quality Control

For our second experiment, we demonstrate the ability of the proposed algorithm to be used in a DTI quality control framework. Here, a collection of 45 subjects consisting of both a T1-weighted image and corresponding DTI images were retrieved from an ongoing study in anonymous from under IRB approval. The T1-weighted images were oriented axially and consisted of $170 \times 256 \times 256$ voxels at 1.0 mm isotropic resolution. The DTI images contained a single B_0 image and 92 diffusion weighted images, with all images consisting of $96 \times 96 \times 52$ voxels and 2.5mm isotropic resolution mm. Due to the difficulty in acquiring consistent and robust DTI images, several of the images within these datasets exhibit problems in terms of image quality (e.g., various degrees of aliasing and shading artifacts).

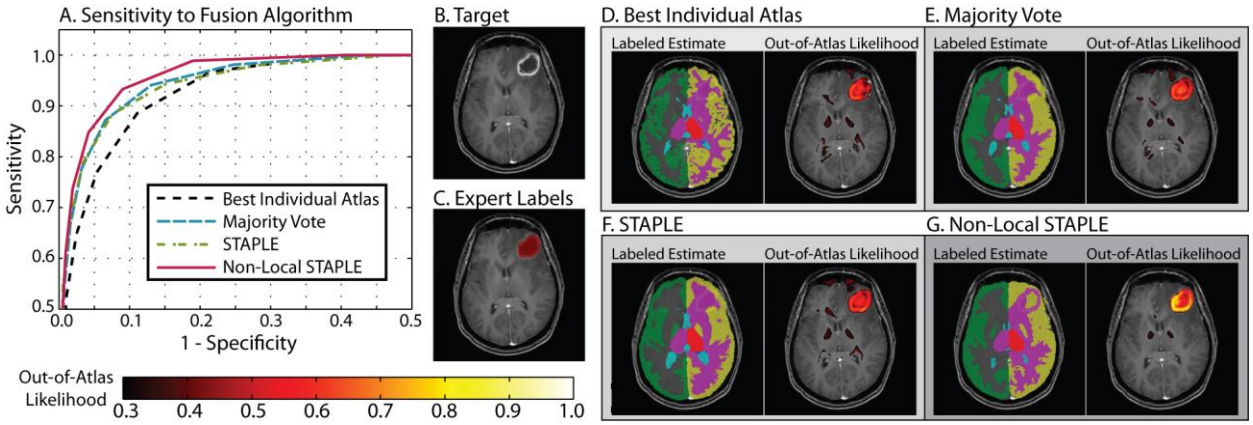


Figure VI.5. Sensitivity of the approach to the label fusion algorithm. A comparison is made between 4 different fusion approaches: (1) best individual atlas, (2) majority vote, (3) STAPLE, and (4) Non-Local STAPLE. Non-Local STAPLE provides both quantitatively and qualitatively the best results due to the fact that it incorporates both label and intensity information into the fusion process. Note that all of the multi-atlas fusion approaches outperform the best individual atlas which highlights the importance of using multiple template images to account for atlas bias.

We employed a two-tier multi-atlas segmentation framework to obtain labels for the DTI images (Figure VI.6). The 15 atlases were registered to the T1-weighted subjects in a pairwise fashion using the SyN non-rigid registration algorithm [36] and the corresponding label observations were fused using NLS. Next, the T1-weighted labels were then transferred to the corresponding B_0 image using an intra-subject rigid registration. Lastly, each of the diffusion weighted images was rigidly registered to their corresponding B_0 image to account for patient movement and to obtain consistent labels for all of the images within each DTI dataset. To assess DTI quality, five of the resulting DTI images were chosen as "atlases" so that the OOA likelihood estimation framework could be applied to the remaining 40 subjects. Note that the B_0 images were normalized to one another using the previously described intensity normalization process and each of the diffusion weighted images were normalized to their corresponding B_0 images in order to obtain consistent intensity values across subjects.

3.5. DTI Results

Qualitative results for this DTI quality control experiment are presented in Figure VI.7. Here, we show 6 representative examples that exhibit varying degrees of image quality issues. The first two examples (the top two rows) represent well controlled datasets and this is supported by the lack of any

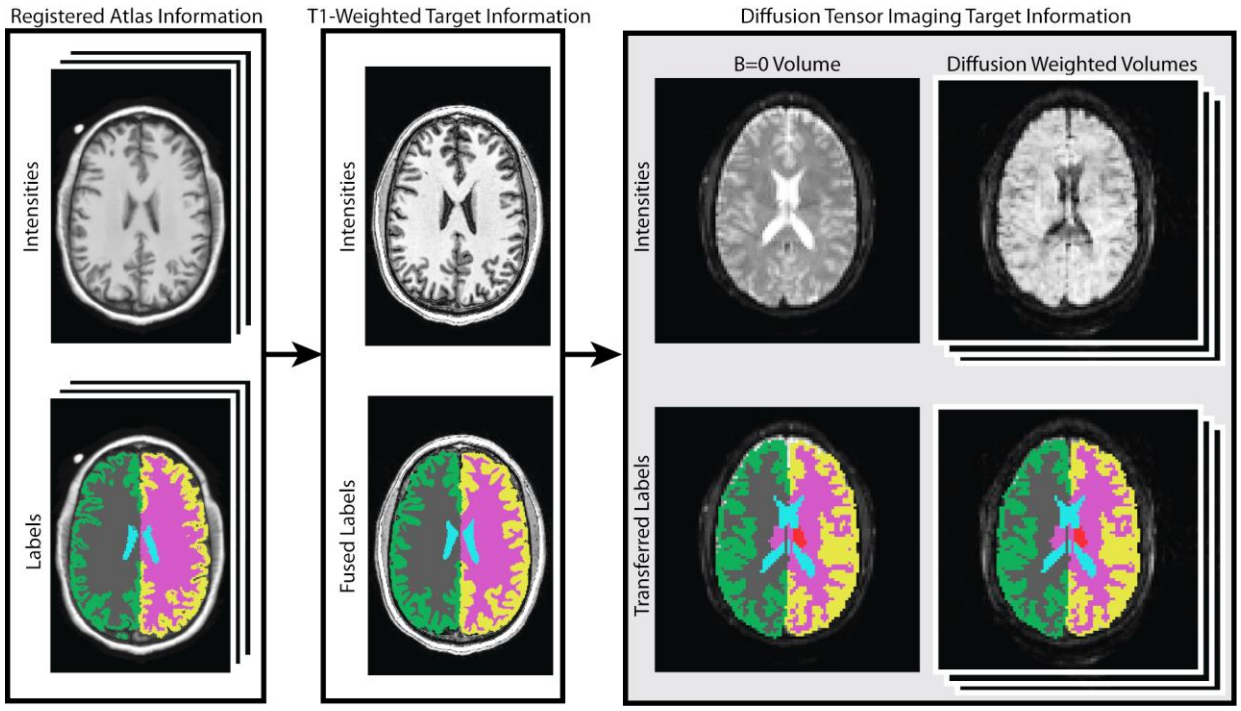


Figure VI.6. Flowchart demonstrating the multi-atlas labeling process for the DTI study. First, the provided atlases are used to label each subject's T1-weighted image. Next, this label information is transferred to all of the DTI datasets via an intra-subject rigid registration. Note that all of the diffusion weighted volumes were rigidly registered to their associated B_0 volume to account for patient movement.

OOA likelihoods greater than 0.3. The example in the third row represents a B_0 image that exhibits an aliasing issue. Here, the OOA likelihood estimate catches this aliasing issue and indicates this anomalous behavior in the appropriate image location. The final three exemplars (the bottom 3 rows) represent diffusion weighted images that exhibit varying degrees of aliasing and shading artifacts. For example, the example in the bottom row represents an example that has severe shading artifacts across more than half of the image. The proposed algorithm clearly detects this large-scale issue and provides consistently high OOA likelihoods across the observed slice.

4. Discussion

The proposed OOA framework extends the multi-atlas labeling paradigm to be sensitive to abnormalities present in the medical images. Previous work on the problem of abnormality detection has primarily relied on a single atlas (or template) [156, 161] and, as a result, has been largely dependent on highly accurate non-rigid registration. Moreover, previous abnormality detection algorithms have been

highly tuned for specific applications (e.g., brain tumor segmentation [156-158, 161], lung nodule detection [159], intestinal abnormalities [160]). The proposed method provides a fully general framework that (1) uses multiple normal atlases to limit the inherent bias of using a single atlas and avoid the need for non-rigid registration, and (2) can be used in a large number of potential applications.

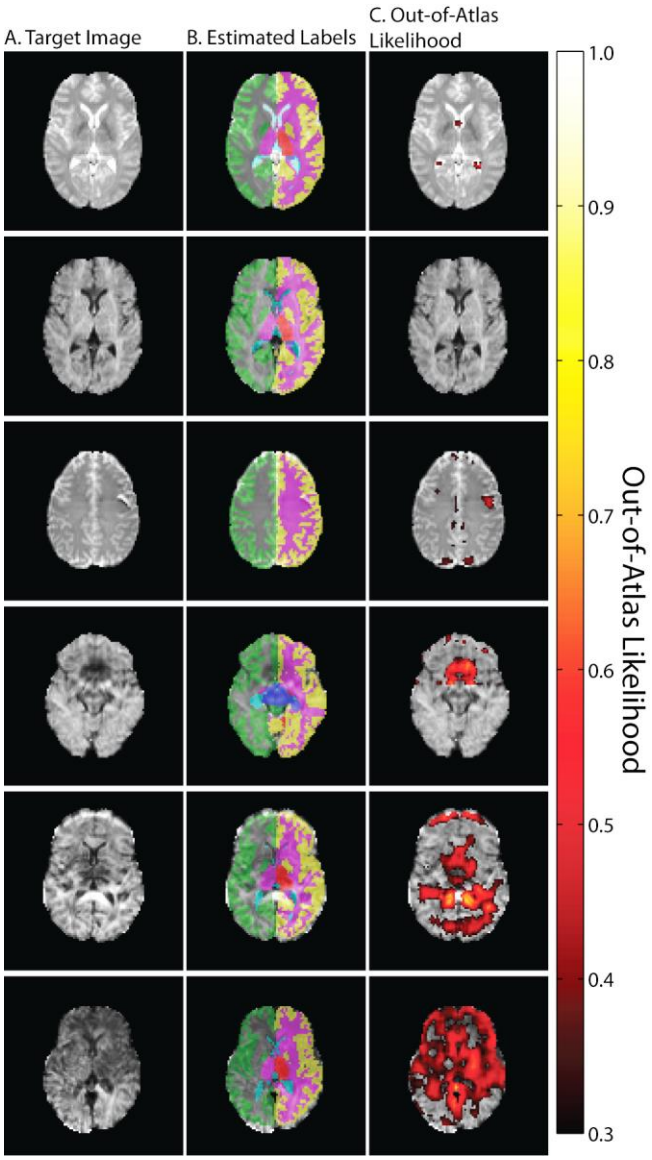


Figure VI.7. Qualitative results for the quality control framework for DTI images. Six representative examples are presented demonstrating the gamut of potential image qualities in the provided dataset. The first two examples (top two rows) represent examples where no abnormalities are present and the out-of-atlas likelihood estimate supports this observation. The final four examples demonstrate images with varying degrees of aliasing and shading artifacts, and the out-of-atlas likelihood estimate consistently detects and localizes these image quality issues.

Despite the promise of the OOA likelihood estimation framework, there are limitations to the proposed approach. First, we use a collection of normal (non-gadolinium enhanced) T1-weighted atlases and use them to assess images that were acquired using clinical imaging protocols (e.g., differing imaging sequence). As a result, the ability to intensity normalize these images is limited and we are forced to limit ourselves to applications where the intensity profile of the desired abnormality is dramatically different than normal anatomy (e.g., malignant gliomas). The use of the proposed framework for the detection of more subtle anatomical pathologies would be inherently limited unless the atlases were constructed using the appropriate imaging characteristics.

Along the same limes, the proposed framework is limited in its ability to detect anomalies that have similar intensity profiles to normal anatomy. For example, differentiating between white matter lesions and gray matter would be difficult using the proposed framework due to the fact that lesions often have similar intensity characteristics as cerebral gray matter. Thus, incorporation of more sophisticated comparison techniques and/or feature vectors would be a promising area of investigation. Direct modeling of texture and shape characteristics into the OOA model, for example, could improve the potential applications by which the model could be applied. Additionally, direct incorporation of label constraints (e.g., topology, symmetry across the cerebral hemispheres) could enable the OOA likelihood estimation framework to use both intensity and label information simultaneously.

In conclusion, the out-of-atlas likelihood estimation framework shows great promise for robust and rapid identification of brain abnormalities and imaging artifacts. Using only weak dependencies on anomaly morphometry and appearance, we demonstrate the ability to (1) detect malignant gliomas on T1-weighted images and (2) identify quality control issues for DTI images. We envision that this approach would allow for application-specific algorithms to focus directly on regions of high OOA likelihood, which would (1) reduce the need for human intervention, and (2) reduce the propensity for false positives. Alternatively, this technique may allow for algorithms to focus on regions of relatively normal anatomy to ascertain image quality or model/adapt to image appearance characteristics.

CHAPTER VII

GROUPWISE SEGMENTATION OF THE SPINAL CORD'S INTERNAL STRUCTURE

1. Overview

The spinal cord is an essential and vulnerable component of the central nervous system which can be significantly affected by numerous neurological conditions – e.g., amyotrophic lateral sclerosis, multiple sclerosis, and neuromyelitis optica [166-170]. Differentiating and localizing pathology/degeneration of the gray matter (GM) and white matter (WM) plays a critical role in assessing the magnitude of tissue damage, therapeutic impacts and determining prognosis of these conditions [171, 172]. While automated methods have been used to segment the spinal cord from the surrounding cerebrospinal fluid (CSF) [173-177] and semi-automated methods have been used for more detailed parcellation of the individual spinal columns [178, 179], automated delineation of internal spinal cord structures (i.e., GM/WM) has not been reported for any imaging modalities. In fact, high-resolution MRI that can provide contrast among spinal cord internal structures has only recently become feasible in clinically acceptable scan times [170, 180-183]. Moreover, MRI of the cervical spinal cord is hindered by numerous technical challenges [184-186] – including (1) the small dimensions (1-2 cm in diameter) with subsequent signal to noise limitations, (2) similarity between WM and GM T1 and T2 values resulting in poor intra-cord contrast [187], (3) involuntary/physiological patient motion, and (4) imaging inhomogeneities and artifacts. Given the challenges associated with spinal anatomy, imaging data, and subsequent processing, developing a robust system to consistently and accurately overcome these challenges is essential. The goal of this manuscript is to provide an efficient and accurate segmentation framework specifically focused on segmenting the spinal cord's internal structure and enabling future clinically relevant inference about the anatomy and its associated conditions.

Over the past decade, multi-atlas segmentation has come to represent the *de facto* standard segmentation framework for its ability to rapidly and accurately generalize structural information from labeled examples (i.e., atlases) [9, 26]. In multi-atlas segmentation, multiple atlases are registered to the target [36, 101] and the resulting voxelwise label conflicts are resolved using label fusion [8, 48, 56, 59, 61, 78]. Since its inception, multi-atlas segmentation has exploded in popularity and has been used across a wide range of potential applications – including, but not limited to, whole-brain [26, 48, 49, 51, 59, 63, 65], hippocampus [56, 61, 118], cardiac [57], prostate [58], and abdomen [75].

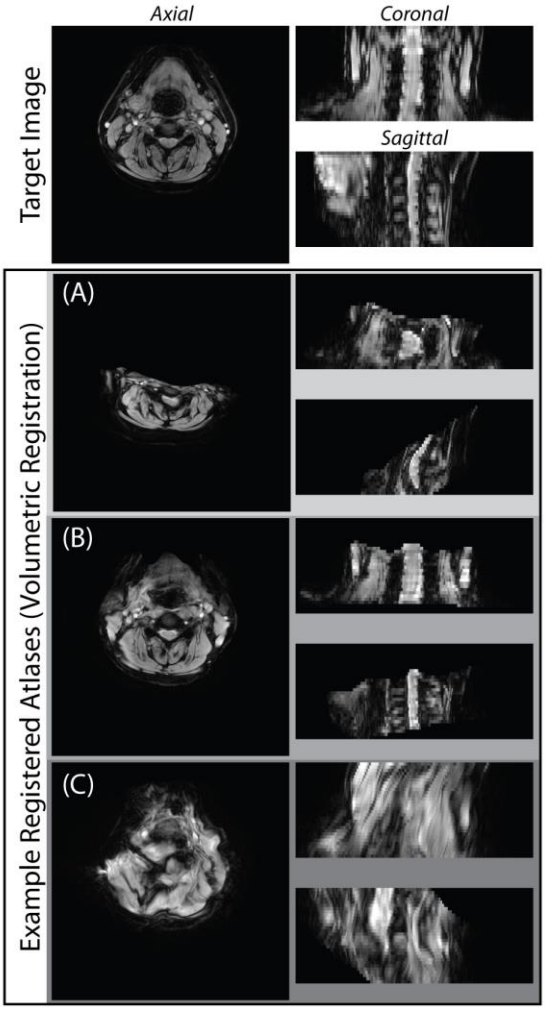


Figure VII.1. Problems associated with non-rigid volumetric registration of cervical spinal cord MRI. Non-rigid volumetric registration of cervical spinal cord MRI is challenging and may yield suboptimal results, including (A) poor global initialization, (B) undesired boundary conditions, and (C) overly smoothed deformations as a result of poor local correspondence.

The manner in which the atlas-target registrations are performed has a substantial impact on multi-atlas segmentation quality [39]. In a typical multi-atlas framework, each atlas is non-rigidly registered to the target in a pairwise fashion (i.e., each atlas-target registration is computed independently). More recently, "groupwise" registration approaches (i.e., pre-alignment of the atlas information to a common groupwise space) have become increasingly popular [65, 105, 106, 109, 188-190]. Groupwise registrations have several benefits, primarily: (1) they reduce the computational burden by requiring only a single registration from the groupwise space to the target space, and (2) through projecting the co-registered atlas information into a low dimensional space (i.e., a "manifold") [65, 188, 191], they provide a natural framework for modeling the relationships between atlases – e.g., for atlas selection [63, 64].

However, regardless of the registration framework, if consistent 3-D deformable registration between the atlases and the target cannot consistently and accurately find structurally-meaningful correspondence, then the applicability of multi-atlas segmentation is unlikely. Unfortunately, due to the extreme variability exhibited among cervical spinal cord 3-D MRI of individual subjects, deformable atlas-target registrations consistently suffer (Figure VII.1) from (a) poor global initialization, (b) undesired boundary conditions, and (c) dramatic over-smoothing due to lack of local correspondence. Comparatively speaking, when viewed from a 2-D axial cross-section within the acquisition plane (i.e., a slice), the inter-subject variability is notably lessened. As a result, when viewed from a slice-based perspective, robustly discovering structurally relevant homology becomes substantially easier. Additionally, a slice-based perspective on the segmentation of cervical MRI (1) enables the use of significantly more (slice-based) atlases, and (2) dramatically reduces the computational burden of individual atlas-target registrations.

Herein, we propose the first approach for fully automated segmentation of cervical spinal cord internal structure using a groupwise slice-based multi-atlas registration framework. Building on the seminal work on "eigenfaces" [192] and active shape/appearance models [105, 106], we provide a method for (1) pre-aligning the slice-based atlas information into a common, groupwise-consistent coordinate

system, (2) constructing a model describing spinal cord variability (i.e., "eigenspines"), (3) registering the target image slice to the model space using a simultaneous intensity- and model-driven cost function, and (4) estimating a final segmentation by fusing the provided atlas information. Additionally, the proposed framework provides a natural mechanism for selecting geodesically appropriate atlases (i.e., atlas selection) and initializing the free model parameters in an informed model-specific context. This presentation is a reformulation of a recently published conference paper [132] in which we provide a completely re-derived groupwise registration framework to provide substantial improvements in overall robustness and accuracy.

In this chapter, we begin by deriving the theoretical framework governing our proposed slice-based groupwise registration framework. Next, we perform a cross-validation experiment (using 67 subjects) in which we demonstrate significant quantitative and qualitative improvement over comparable volumetric and slice-based pairwise registration frameworks (in terms of GM and WM segmentation accuracy). Lastly, the sensitivity of the proposed model is addressed with respect to the free model parameters.

2. Theory

2.1. Problem Definition

Consider a target gray-level image represented as a vector, $\boldsymbol{I} \in \mathbb{R}^{N\times 1}$, where N is the number of voxels in the image. Let $\boldsymbol{T} \in \boldsymbol{L}^{N\times 1}$ be the latent representation of the true target segmentation, where $\boldsymbol{L} = \{0, \dots, L-1\}$ is the set of possible labels that can be assigned to a given voxel. Consider a collection of J atlases with associated intensity values, $\boldsymbol{A} \in \mathbb{R}^{N\times J}$, and label decisions, $\boldsymbol{D} \in \boldsymbol{L}^{N\times J}$. The index variables i, j, and l will be used to iterate over the voxels, atlases, and labels, respectively. To summarize, the goal of the proposed approach is to:

- Rigidly register all of the atlas information, $\{\boldsymbol{A}, \boldsymbol{D}\}$, to a common groupwise space. In other words, we need to construct $\{\boldsymbol{A}^M \in \mathbb{R}^{N_M\times J}, \boldsymbol{D}^M \in \boldsymbol{L}^{N_M\times J}\}$, where $N_M < N$ is the number of

voxels that are consistent among the atlases in the common model space (i.e., the relevant subset cropped from each slice).

- Construct the appearance model using the groupwise-consistent atlas representations. The appearance model consists of (1) the mean image, $\boldsymbol{\Psi} \in \mathbb{R}^{N_M \times 1}$(2) the orthogonal eigenvectors ("eigenspines") describing the modes of variation, $\boldsymbol{u} \in \mathbb{R}^{N_M \times V}$, (3) the associated eigenvalues, $\boldsymbol{\lambda} \in \mathbb{R}^{V \times 1}$, and (4) the weights associated with each atlas when projected into the model space, $\boldsymbol{\omega}^M \in \mathbb{R}^{V \times J}$. Note, $V < J \ll N_M$ represents the number of modes of variation that are used in the model.
- Find the optimal rigid transformation, $\boldsymbol{R}$, that maps the target image, $\boldsymbol{I}$, to the groupwise model space, $\boldsymbol{I}^M = \boldsymbol{R} \circ \boldsymbol{I}$, by minimizing a model-driven cost function. Project $\boldsymbol{I}^M$ into the low-dimensional model space, described by $\boldsymbol{\omega} \in \mathbb{R}^{V \times 1}$.
- Using $\boldsymbol{\omega}$ and $\boldsymbol{\omega}^M$, select the geodesically-closest $K \leq J$ atlases to use. Label fusion is then used to fuse the selected K atlases resulting in a segmentation estimate of $\boldsymbol{I}^M$: $\widehat{\boldsymbol{T}}^M \in \boldsymbol{L}^{N_M \times 1}$,.
- Transform $\widehat{\boldsymbol{T}}^M$ to $\widehat{\boldsymbol{T}} \in \boldsymbol{L}^{N \times 1} = \boldsymbol{R}^{-1} \circ \widehat{\boldsymbol{T}}^M$, an estimate of the desired segmentation in the original target space.

Note, as the model we are constructing describes a 2-D (slice-based) representation of the spinal cord, all rigid transformations described in this manuscript are 2-D three degree-of-freedom transformations. In other words, a given transformation matrix, $\boldsymbol{R}$, can be described by $\{t_x, t_y, \theta\}$, where $\{t_x, t_y\}$ represent a two-dimensional translation, and θ represents a rotation angle. We will use the notation, $\boldsymbol{R} \circ \boldsymbol{X}$, to denote the application of transformation matrix $\boldsymbol{R}$ to image $\boldsymbol{X}$.

2.2. Creation of a Groupwise Consistent Atlas Representation

The first step in constructing the appearance model describing the slice-based representations of the spinal cord is to register the atlas information, $\{\boldsymbol{A}, \boldsymbol{D}\}$, to a common groupwise space (i.e., a space in

which all atlas slices are co-aligned). Here, we use an iterative process to construct the groupwise consistent representation of the atlases.

Let $\boldsymbol{\chi}^{(k)} \in \boldsymbol{L}^{N_M \times 1}$ represent the "mean" segmentation image (herein, via majority vote – see eq. 2) at the k^{th} iteration of the described procedure. Then, for each individual atlas, $\{\boldsymbol{A_j}, \boldsymbol{D_j}\}$ the goal is to optimize the following label-based cost function:

$$\begin{aligned} \boldsymbol{R}_j' &= \arg\min_{\boldsymbol{R}_j'} \left\| \boldsymbol{\chi}^{(k)} - \boldsymbol{R}_j' \circ \boldsymbol{D}_j \right\|_0 \\ &= \arg\min_{\boldsymbol{R}_j'} \left\| \boldsymbol{\chi}^{(k)} - \boldsymbol{D}_j^M \right\|_0 \end{aligned} \tag{7.1}$$

where $\boldsymbol{R}_j'$ is the rigid transformation that minimizes the L^0 norm (i.e., the discrete metric) between the current estimate of the groupwise mean segmentation, $\boldsymbol{\chi}^{(k)}$, and the j^{th} atlas segmentation transformed into the current model space, $\boldsymbol{D}_j^M = \boldsymbol{R}_j^{'} \circ \boldsymbol{D}_j$.

After the optimized rigid transformations are computed for each individual atlas, the mean segmentation estimate can be updated using majority vote fusion:

$$\boldsymbol{\chi}_i^{(k+1)} = \arg\max_{l \in \boldsymbol{L}} \sum_j \delta\left(D_{ij}^M, l\right) \tag{7.2}$$

where $\delta(\cdot,\cdot)$ is the Kronecker delta function. The process defined by (1) and (2) is then iterated until the mean segmentation estimate converges to a consistent segmentation (i.e., no change in the L^0 norm between successive iterations: $\left\| \boldsymbol{\chi}^{(k-1)} - \boldsymbol{\chi}^{(k)} \right\|_0 = 0$). In practice, convergence typically occurred in fewer than 5 iterations. Note, an alternative approach could be to use Expectation-Maximization (EM) to fuse the initial registered atlases for construction of the groupwise consistent mean (e.g., [193]); however, unlike the proposed method, this type of framework would largely rely on the success of the initial registrations. Lastly, the groupwise consistent representation of the atlases (i.e., $\{\boldsymbol{A}^M \in \mathbb{R}^{N_M \times J}, \boldsymbol{D}^M \in \boldsymbol{L}^{N_M \times J}\}$) is then constructed using the final transformations, $\boldsymbol{R}_j'$, after the iterative registration procedure has converged:

$$\left\{\boldsymbol{A}_j^M, \boldsymbol{D}_j^M\right\} = \{\boldsymbol{R}_j' \circ \boldsymbol{A}_j, \boldsymbol{R}_j' \circ \boldsymbol{D}_j\} \tag{7.3}$$

2.3. Appearance Model Construction

Now that we have all of the atlas information in a groupwise consistent representation, $\{\boldsymbol{A}^M, \boldsymbol{D}^M\}$, it is possible to construct an appearance model. The first step is to find the mean atlas image, $\boldsymbol{\Psi} \in \mathbb{R}^{N_M \times 1}$ using the aligned atlas image information

$$\boldsymbol{\Psi} = \frac{1}{J}\sum_j \boldsymbol{A}_j^M. \tag{7.4}$$

Next, we find the eigenvectors (referred to as "eigenspines"), $\boldsymbol{u} \in \mathbb{R}^{N_M \times V}$, and eigenvalues, $\boldsymbol{\lambda} \in \mathbb{R}^{V \times 1}$ using principal component analysis (PCA) [194]. As a result, for the v^{th} eigenspine/eigenvalue pair, $\boldsymbol{u}_v$ is chosen such that the following function is maximized

$$\lambda_v = \frac{1}{J}\sum_j \left[\boldsymbol{u}_v^T\left(\boldsymbol{A}_j^M - \boldsymbol{\Psi}\right)\right]^2 \tag{7.5}$$

subject to the constraint that all eigenspines are unit vectors (i.e., $|\boldsymbol{u}_v| = 1$) and orthogonal (i.e., $\boldsymbol{u}_v{}^T\boldsymbol{u}_{v'} = \delta(v, v')\ \forall v \neq v'$). Note that the number of eigenspine/eigenvalue pairs that are constructed is a function of the desired fraction of the total variance that the model should explain, κ. This number, V, is chosen as the minimum value for which

$$\frac{\sum_{v=1}^{V}\lambda_v}{\sum_{v'=1}^{J}\lambda_{v'}} > \kappa \tag{7.6}$$

Finally, we compute the projection weights for each of the atlases that were used in constructing the model. These weights, $\boldsymbol{\omega}^M \in \mathbb{R}^{V \times J}$, where each element, $\boldsymbol{\omega}_j^M \in \mathbb{R}^{V \times 1}$ is found through a vector projection:

$$\boldsymbol{\omega}_j^M = \boldsymbol{u}^T\left(\boldsymbol{A}_j^M - \boldsymbol{\Psi}\right) \tag{7.7}$$

Thus, $\boldsymbol{\omega}_j^M$ can be directly interpreted as the relative amount of variance explained by each of the V available modes of variation, for a given atlas j.

As an important aside, in standard practical applications, the creation of the groupwise consistent atlas representation (eqs. 1-3) and the construction of the appearance model (eqs. 4-7) would be

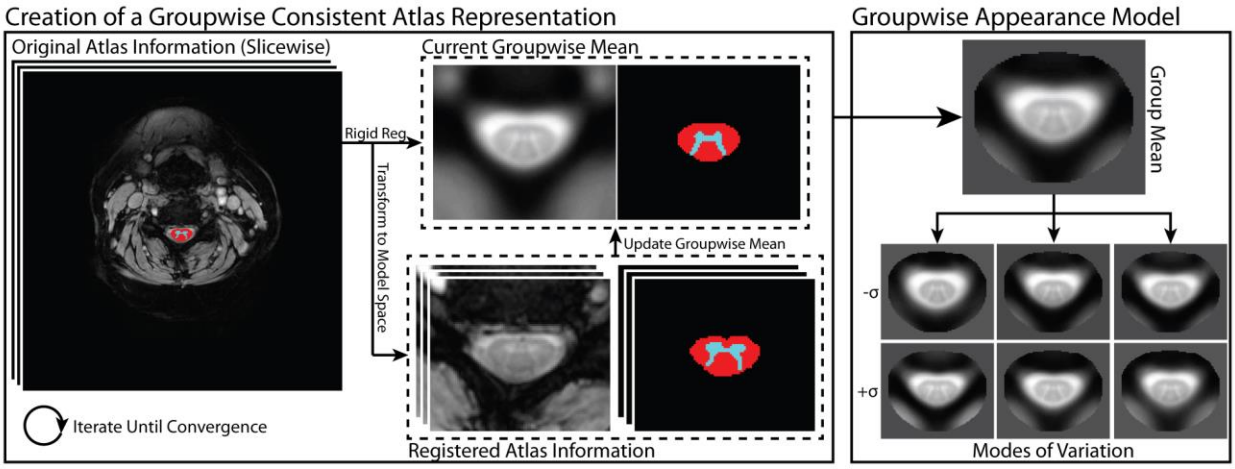


Figure VII.2. Flowchart describing the construction of the groupwise consistent atlas representation and the resulting groupwise appearance model. In an iterative procedure, all of the atlases registered to the current estimate of the mean, which is then updated. Using the co-registered atlas data, the groupwise appearance model is constructed using principal component analysis.

performed entirely "offline" (i.e., a new model would not need to be constructed for each desired target volume).

Figure VII.2 provides a visual representation of the theoretically described process of both creating the groupwise-consistent atlas representations and, subsequently, understanding the resulting appearance model. Figure VII.3 visualizes local relationships between the atlases when they are projected into the low-dimensional appearance model space, and demonstrates that the low dimensional representation maintains the relationships between the visually expected atlas similarities.

2.4. Groupwise Registration and Segmentation using the Appearance Model

Once the appearance model for the atlases has been constructed, the primary remaining challenge is to find the rigid transformation, $\boldsymbol{R}$, that maps each target image, $\boldsymbol{I}$, into the model space – represented by $\boldsymbol{I}^M$. We define a cost function that is a function of both (1) how well the model represents $\boldsymbol{I}^M$, and (2) the image similarity between $\boldsymbol{I}^M$ and a weighted average representation of the atlases, $\hat{\boldsymbol{I}}^M$. Thus, for a given transformation, $\boldsymbol{R}$, the projection weights for the target image, $\boldsymbol{\omega} \in \mathbb{R}^{V\times 1}$ can be found similarly to (7) as

$$\begin{aligned}\boldsymbol{\omega} &= \boldsymbol{u}^T(\boldsymbol{R} \circ \boldsymbol{I} - \boldsymbol{\Psi}) \\ &= \boldsymbol{u}^T(\boldsymbol{I}^M - \boldsymbol{\Psi})\end{aligned} \tag{7.8}$$

We can then define the model similarity between all of the individual atlases and the projected target image, $\boldsymbol{\beta} \in \mathbb{R}^{J\times 1}$, using the current estimate of the desired transformation matrix, $\boldsymbol{R}$. Let β_j (i.e., the similarity between the j^{th} atlas and the projected target image) be defined as

$$\beta_j = \frac{1}{\mathrm{Z}} \exp\left(-\tau \left\| \boldsymbol{\omega} - \boldsymbol{\omega}_j^M \right\|_2\right) \tag{7.9}$$

where Z is the partition function which enforces the constraint that $\sum_j \beta_j = 1$, and τ is a model weighting parameter indicating the decay constant associated with the geodesic distance between a given atlas and the projected target image in model space. Using these similarity weights, we construct (1) the weighted mean of the atlas projection weights, $\boldsymbol{\mu} \in \mathbb{R}^{V\times 1}$, (2) the weighted standard deviation of the projection weights $\boldsymbol{\sigma} \in \mathbb{R}^{V\times 1}$, and (3) the weighted image representation of the projected target image from the

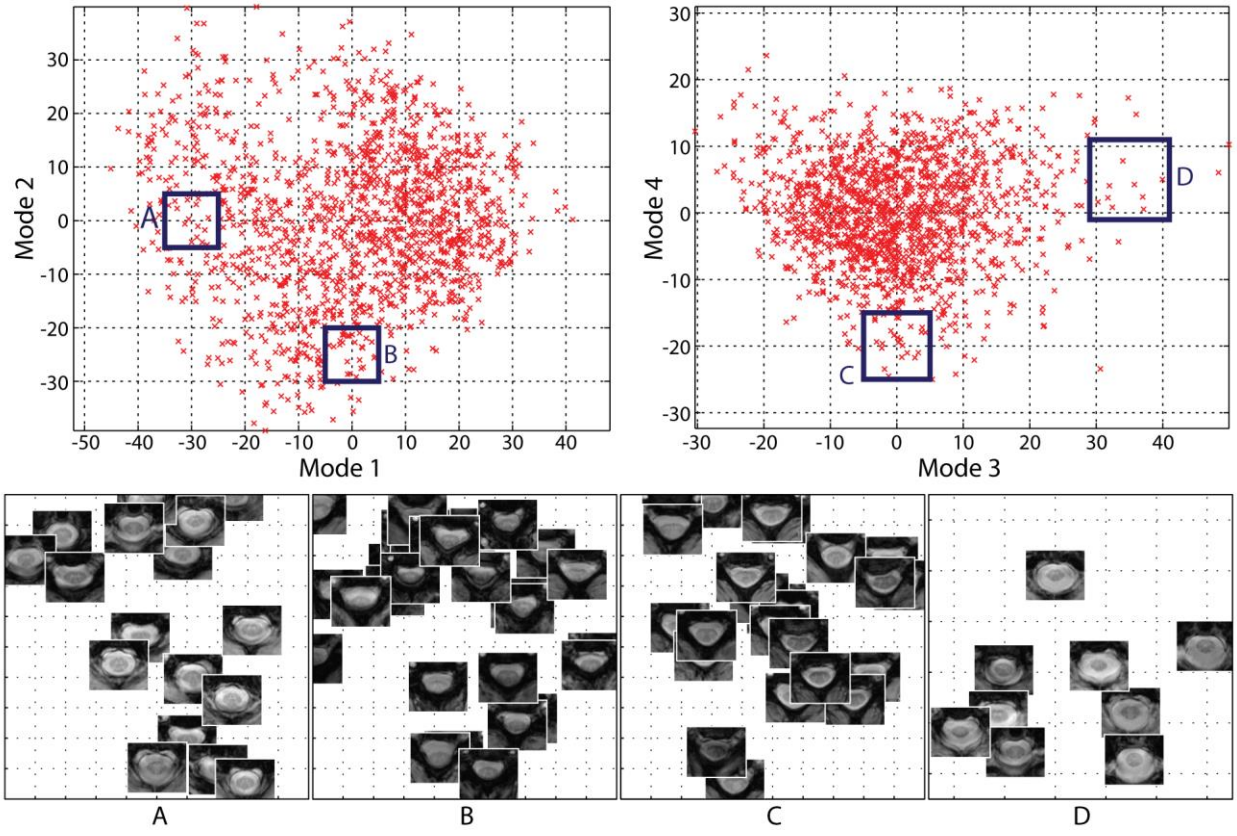


Figure VII.3. Example local atlas image content with respect to the primary modes of variation constructed in the model. The geodesic distance between co-registered atlases (i.e., the distance in the low-dimensional model) visually corresponds to anatomical similarity.

atlases, $\hat{\boldsymbol{I}}^M \in \mathbb{R}^{N_M \times 1}$, where

$$\boldsymbol{\mu} = \boldsymbol{\omega}^M \boldsymbol{\beta} \tag{7.10}$$

$$\sigma_v = \sum_j \beta_j \left(\omega_{vj}^M - \mu_v\right)^2, \forall\, v \in \{1 \dots V\} \tag{7.11}$$

$$\hat{\boldsymbol{I}}^M = \boldsymbol{A}^M \boldsymbol{\beta} \tag{7.12}$$

Using (10)-(12), we define the registration cost function for aligning the target image to the model space, $\boldsymbol{R}$, to be

$$\begin{aligned} \boldsymbol{R} &= \arg\min_{\boldsymbol{R}} \left(\left\|\hat{\boldsymbol{I}}^M - \boldsymbol{R} \circ \boldsymbol{I}\right\|_2^2\right)\left(\sum_v \frac{(\omega_v - \mu_v)^2}{\sigma_v^2}\right) \\ &= \arg\min_{\boldsymbol{R}} \left(\left\|\hat{\boldsymbol{I}}^M - \boldsymbol{I}^M\right\|_2^2\right)\left(\sum_v \frac{(\omega_v - \mu_v)^2}{\sigma_v^2}\right) \end{aligned} \tag{7.13}$$

where the first term, $\left(\left\|\hat{\boldsymbol{I}}^M - \boldsymbol{I}^M\right\|_2^2\right)$, represents the squared L^2 norm between the projected target image and the weighted average representation of the projected target image (i.e., using the model weights and the atlas image information), and the second term, $\left(\sum_v \frac{(\omega_v - \mu_v)^2}{\sigma_v^2}\right)$ approximates the Gaussian log-likelihood of the observed weights of the projected target image, $\boldsymbol{\omega}$, given the provided appearance model.

Given the optimal rigid registration, $\boldsymbol{R}$, that maps the target image into the appearance model space, it is possible to estimate the underlying target segmentation using label fusion. The geodesic distance-based similarity weights, $\boldsymbol{\beta}$, provide a natural and straightforward mechanism for performing informed atlas selection [63, 64] – a claim qualitatively supported in Figure VII.3. Here, we select the set of all atlases for which β_j is greater than some arbitrary constant, ϵ. Let $\left\{\boldsymbol{A}_{\boldsymbol{K}}^M, \boldsymbol{D}_{\boldsymbol{K}}^M\right\}$ be the set of selected atlases, where $\boldsymbol{K} = \{k | \beta_k > \epsilon\}$. Thus, the segmentation estimate of the target image in model space, $\hat{\boldsymbol{T}}^M$, is

$$\hat{T}_i^M = \arg\max_{l \in \boldsymbol{L}} f\left(\hat{T}_i^M = l \middle| \left\{\boldsymbol{A}_{\boldsymbol{K}}^M, \boldsymbol{D}_{\boldsymbol{K}}^M\right\}, \boldsymbol{I}^M\right) \tag{7.14}$$

where $f\left(\hat{T}_i^M = l \middle| \left\{\boldsymbol{A}_{\boldsymbol{K}}^M, \boldsymbol{D}_{\boldsymbol{K}}^M\right\}, \boldsymbol{I}^M\right)$ can be approximated using a pre-defined label fusion framework – e.g., [8, 26, 51, 59, 61, 78].

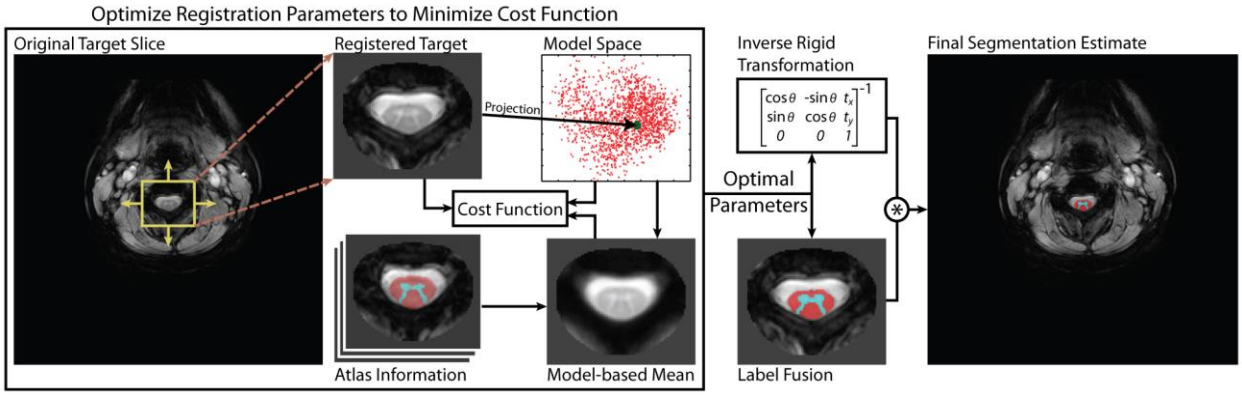


Figure VII.4. Flowchart describing the process of (1) registering the target image with the model space, and (2) constructing the final segmentation estimate. For a given rigid transform, the registered target is projected into the model space enable a model-informed cost function to be evaluated. This process is repeated until the optimal rigid transformation is found. Using the optimal parameters (i.e., the transform and the selected atlas content), the segmentation estimate is constructed through label fusion and, finally, transferred back to the original target coordinate system.

Finally, the last remaining step is to transfer the segmentation estimate from model space, $\widehat{\boldsymbol{T}}^{M}$, to the original space defined by the target image $\widehat{\boldsymbol{T}}$. This can be easily accomplished using the inverse of the optimal rigid transformation found in (13).

$$\widehat{\boldsymbol{T}} = \boldsymbol{R}^{-1} \circ \widehat{\boldsymbol{T}}^{M} \tag{7.15}$$

Figure VII.4 provides a visual representation of the described framework for both registering and projecting the target image into the low-dimensional model space, and constructing and transferring the label fusion estimate into the original target coordinate system.

2.5. Model Parameters and Initialization

The proposed model has three primary parameters:

1. The fraction of the explained model variance to use, κ, and, thus, the number of modes of variation to use, V (eq. 6).
2. The weighting parameter indicating the decay constant associated with the geodesic distance between a given atlas and the projected target image in model space, τ (eq. 9).
3. The threshold for determining which atlases to keep when performing the label fusion step, ϵ (eq. 14).

First, the fraction of the explained model variance to use, κ, and, thus, the number of modes of variation to use, V, plays a critical role in both registration and segmentation accuracy. Sub-optimal values for this parameter can have adverse effects on the quality of the registration and segmentation. A value that is too small (i.e., using too few modes of variation) can result in an inability to accurately represent the variability exhibited within the target image. As a result, the projection of the registered target image could indicate erroneous relationships with the provided atlases. Alternatively, values that are too large (i.e., using too many modes of variation) could lead to the inclusion of modes of variation that are simply noise (i.e., modes that do not provide meaningful information on the variability of the cervical spinal cord).

Second, the weighting parameter, τ (eq. 9), provides a mechanism for converting the geodesic distance between each individual atlas and the projected target image into relative weights for each atlas. This parameter can be viewed as a proxy for determining the influence of geodesic distance within the low-dimensional model space. To illustrate, for values of τ that are too *small*, the influence of atlases that are far away from the projected target image will be given too much influence on the registration/segmentation framework. Alternatively, for values of τ that are too *large*, only the closest atlases to the projected target image will be given any substantial influence, resulting in a representation of the target image that is too sparse. Fortunately, the estimated model provides a natural mechanism for *a priori* estimation of this parameter. For ease of representation, let $\boldsymbol{K}_j = \left\{ j' \neq j \middle| \exp\left(-\tau \left\| \boldsymbol{\omega}_j^M - \boldsymbol{\omega}_{j'}^M \right\|_2\right) > \epsilon \right\}$ represent the collection of atlases that would be chosen for atlas j given the model parameters $\{\tau, \epsilon\}$. For a fixed ϵ, the optimal value for τ can be approximated as:

$$\begin{aligned} \tau_{est} &= \arg\min_{\boldsymbol{\tau}} \sum_j \sum_i \left\| D_{ij}^M - \arg\max_{l \in \boldsymbol{L}} \sum_{j' \in \boldsymbol{K}_j} \delta\left(D_{ij'}^M, l\right) \right\|_0 \\ &= \arg\min_{\boldsymbol{\tau}} \sum_j \left\| \boldsymbol{D}_j^M - \widehat{\boldsymbol{D}}_j^M \right\|_0 \end{aligned} \tag{7.16}$$

where $\widehat{\boldsymbol{D}}_j^M = \arg\max_{l \in \boldsymbol{L}} \sum_{j' \in \boldsymbol{K}_j} \sum_i \delta\left(D_{ij'}^M, l\right)$ is an estimate (via a majority vote) of the j^{th} atlas segmentation of voxel i using a given parameter value, τ, and the remaining $J-1$ atlas segmentations. In other words, we choose τ_{est} such that the resulting estimated atlas segmentations are closest to the original, known, atlas segmentations.

Lastly, observe that the threshold for determining which atlases to keep when performing the label fusion step, ϵ, is very tightly related to the value chosen for the weighting parameter, τ. In eq. 16, we fixed the value of ϵ in order to directly optimize τ. Empirically, we found that for a given change in ϵ, the optimal value for τ would change accordingly (i.e., an increase/decrease in ϵ would cause a corresponding increase/decrease in τ) without affecting the resulting segmentation accuracy. As a result of this direct relationship, we use a constant value of $\epsilon = 10^{-3}$ for all subsequent experiments.

3. Methods and Results

3.1. Data

Herein, we study a dataset consisting of 67 MR volumes of the cervical spinal cord. All data were acquired axially on a 3T Philips Achieva scanner (Philips Medical Systems, Best, The Netherlands) using a single channel body coil for transmission and a 16-channel neurovascular coil for signal reception. The center of the imaging volume was aligned to the space between the 3rd and 4th cervical vertebrae. T2*-weighted data were obtained using a 3D gradient echo (TR/TE/a = 121/12ms/9°) with a multi-shot EPI (EPI factor = 3) covering a field of view of 190 x 224 x 90 mm^3 with nominal resolution of 0.6 x 0.6 x 3 mm^3. Fat saturation was implemented using a 1331 binomial water excitation (ProSet), 2 signal averages, and a SENSE factor of 2.

The "gold standard" manual labels were constructed by an experienced rater who is familiar with MR images of the cervical spinal cord. The labeling process was performed using the Medical Image Processing, Analysis and Visualization (MIPAV) software [165]. For each slice, two labels were considered: the white matter (WM) and the gray matter (GM) (often referred to as the gray matter horns).

Due to imaging artifacts and lack of reasonable contrast, not all slices on each volume were labeled (in total, a collection of 1538 slices were labeled). As a result, all volumetric accuracy measurements were constructed using only the slices that were labeled by the experienced rater and all un-labeled slices were ignored.

3.2. Implementation of Proposed Framework

All computations for the proposed groupwise registration framework were performed using MATLAB (Mathworks, Natick, MA) on a 64-bit quad-core 3.07 GHz desktop computer with 13GB of RAM, running Ubuntu 12.04. The minimization/maximization procedures found in eqs. 1, 13 and 16 were coarsely initialized using a line search and finely optimized using the Nelder-Mead simplex direct search algorithm [195]. Additionally, for registering the target image to model space (eq. 13), the rigid transformations were initialized such that they aligned the center of mass of the target image with the model space. To eliminate undue impact of uninformative voxels, all image representations in the derived model space were cropped to include all voxels that were within 10mm of the final mean segmentation of the groupwise-consistent slice-based atlases (forming N_m in eq. 2) obtained from the model (see Figure VII.2). While this value could certainly be considered an additional model parameter, we empirically found that, as long as we ignored most of the background information, the exact amount of cropping did not noticeably impact registration/segmentation accuracy. Lastly, to simplify the direct comparison of image intensities (e.g., eq. 13), the registered target image and all of the atlas images were normalized to a unit Gaussian distribution within the region-of-interest specified by the defined model space.

3.3. Baseline Approaches

3.3.1. Registration

To assess the accuracy of the proposed framework we consider three registration procedures: (1) a pairwise volumetric non-rigid registration framework, (2) a pairwise slice-based rigid registration framework, and, (3) the proposed groupwise slice-based rigid registration framework – as described

above. The procedures for the baseline pairwise volumetric and slice-based registration frameworks are described below.

For the pairwise volumetric non-rigid registration, we used the SyN non-rigid registration algorithm [36] (as part of the Advanced Normalization Tools [ANTs] toolkit, http://stnava.github.io/ANTs/). In recent comparisons [39, 196], the SyN registration algorithm has been consistently shown to be the premier registration algorithm for multi-atlas segmentation. For all presented results, we used the default parameters for large deformation mapping described in [197].

For the pairwise slice-based rigid registration, we used the symmetric rigid registration algorithm based on [101] (available as "reg_aladin" in the "NiftyReg" registration package -- http://sourceforge.net/projects/niftyreg/). Due to the symmetric constraint on the resulting registration, we have empirically found this method to consistently result in fewer failures when compared to alternative unidirectional rigid registration approaches – e.g., [98]. The default parameters provided by the NiftyReg software package were used for all presented results. Note, for fairness of comparison, atlas images in the slice-based pairwise registration procedure were cropped in the same manner as the described groupwise registration framework to minimize failures due to the inclusion of uninformative background information.

3.3.2. Label Fusion

As we are interested in multi-atlas segmentation, the label fusion technique used to combine (or "fuse") the resulting registered atlases plays a critical role in segmentation accuracy. Herein, we consider a variety of label fusion algorithms that represent the gamut of the state-of-the-art label fusion approaches. The label fusion approaches we consider are: (1) majority vote (MV) [26, 59], (2) locally weighted vote (LWV) [48, 57, 59, 61], (3) patch-based segmentation (PBS) [56], (4) Simultaneous Truth and Performance Level Estimation (STAPLE) [8, 9], and (5) Non-Local STAPLE (NLS) [51, 78].

For MV and LWV, the results were obtained using the exact same approach as described in [59]. For STAPLE, convergence of the Expectation-Maximization (EM) [116] framework was detected when

the average change in the trace of the performance-level parameters fell below 10^{-4}, "consensus voxels" (i.e., voxels where all raters agree) were ignored in the estimation process, and a voxelwise label prior (governed by the result of a probabilistic majority vote) was used. For PBS, a 1mm isotropic search neighborhood and 2mm isotropic patch neighborhood were used. The remaining parameters were chosen in exactly the same manner as described in [56]. Finally, where applicable, NLS used the same initialization parameters as described above for STAPLE and the same non-local correspondence parameters as described for PBS. Additionally, for NLS, the locally normalized correlation coefficient (LNCC) [48, 118] was used for the weighting in the non-local correspondence model – see [132] for details.

The implementation of all label fusion algorithms presented in this manuscript are available as part of the Java Image Science Toolkit (JIST, www.nitrc.org/projects/jist) [152].

3.4. Experimental Methods

Herein, we present two experiments to demonstrate the benefits of the proposed slice-based groupwise registration framework.

3.4.1. Leave-One-Out Cross-Validation

For the first experiment, we perform a leave-one-out cross-validation (LOOCV) to demonstrate the effects of (1) the three considered registration frameworks, and (2) label fusion accuracy for each registration framework. For the pairwise volumetric registration, we non-rigidly registered all 66 non-target volumes to the target volume independently. For the pairwise slice-based registration, due to practical limitations in registering over 1500 slices to each target slice, we chose the 85 slice atlases (excluding the slices from the target) that exhibited the largest GM/WM contrast (difference in intensity profiles) – measured using the Kullback-Leibler divergence [164] between kernel density estimates of the intensity values associated with each structure. Finally, for the proposed groupwise slice-based registration framework, the groupwise appearance model was constructed uniquely for each target volume (i.e., the slices in the target volume were not used in the construction of the model). Note, for standard use

(i.e., situations with a well-defined labeled training set and an unlabeled testing set), the model would be identical for all target volumes in the testing set and, thus, could be computed entirely offline. For all presented experiments, the fraction of explained model variance, κ, was set to 0.99, the model weighting parameter, τ, was set to τ_{est} (see Eq. 16), and the atlas selection threshold, ϵ, was set to 10^{-3}. For reference, using these values for $\{\tau, \epsilon\}$ resulted in the use of, on average, 460 modes of variation and 25 atlases per target image slice.

For each registration framework, the accuracy of the resulting segmentations are presented both quantitatively (Figure VII.5) and qualitatively (Figures VII.6 and VII.7). For the quantitative results, the volumetric accuracy of the resulting segmentations when compared to the manual labels for both gray matter and white matter are considered using: (1) the Dice similarity coefficient (DSC) [140], (2) the symmetric (or bi-directional) mean surface distance error (MSDE), and (3) the symmetric (or bi-directional) Hausdorff distance error (HDE) [149]. For the qualitative results, we show representative example segmentations from both a slice-based perspective (Figure VII.6) and volumetric perspective (Figure VII.7). Due to the consistently poor performance of the typical pairwise volumetric registration framework (e.g., Figure VII.1), only the results of a majority vote are considered in the quantitative results (Figure VII.5). Additionally, this approach is excluded in the qualitative comparison (Figures VII.6 and VII.7) to simplify the visual comparison.

3.4.2. Sensitivity to Model Parameters

For the second experiment, we assess the accuracy of the proposed groupwise slice-based registration framework with respect to the free model parameters. As discussed above, there are two primary parameters that need to be assessed: (1) the fraction of the explained model variance to use, κ, and, thus, the number of modes of variation to use, V (see eq. 6) and (2) the weighting parameter indicating the decay constant associated with the geodesic distance between a given atlas and the projected target image in model space, τ (see eq. 9). For κ, eight unique values are considered ranging from $\kappa = 0.7$ to $\kappa = 0.9999$, while holding the weighting parameter constant, $\tau = \tau_{est}$ (Figure VII.8A).

For τ, 21 unique values are considered in reference to the estimated parameter value τ_{est} (Eq. 16) ranging from $\tau - \tau_{est} = -0.2$ to $\tau - \tau_{est} = 0.2$, while holding the explained model variance constant, $\kappa = 0.99$, (Figure VII.8B). For both parameters, the accuracy of the resulting segmentations is assessed using the mean DSC across both labels. Lastly, in order to isolate the impact of the considered parameter, the results are assessed using a simple majority vote fusion – alternative fusion approaches would obfuscate the impact of the considered parameter with the impact of the more sophisticated fusion approaches.

3.5. Experimental Results

3.5.1. Leave-One-Out Cross-Validation

The quantitative results of the LOOCV experiment (Figure VII.5) demonstrate consistent improvement in overall segmentation accuracy for each of the considered metrics. First, due to the inherent limitations of the traditional pairwise volumetric registration framework when applied to the spinal cord, it is immediately evident that the volumetric approach is consistently outperformed by both of the slice-based approaches for both WM and GM across all of the considered metrics. In particular, due to the poor initialization and the resulting boundary errors, the volumetric approach performs considerably worse than the slice-based approaches in terms of the resulting HDE (the median HDE is approximately 5mm worse than the pairwise slice-based registration and close to 8mm worse than the proposed groupwise registration).

Second, the proposed groupwise slice-based registration framework provides substantial improvement in terms of both robustness and accuracy over the pairwise slice-based registration framework. For a given label fusion algorithm, the segmentations using the groupwise framework significantly outperformed its pairwise counterpart (paired t-test, $p < 0.01$) for both GM and WM across each of the considered accuracy metrics. In particular, the distance-based metrics demonstrate substantial improvements in terms of overall robustness provided by the proposed approach. For the worst case, the groupwise framework provides an improvement of approximately 5mm and 8.5mm in MSDE for the GM

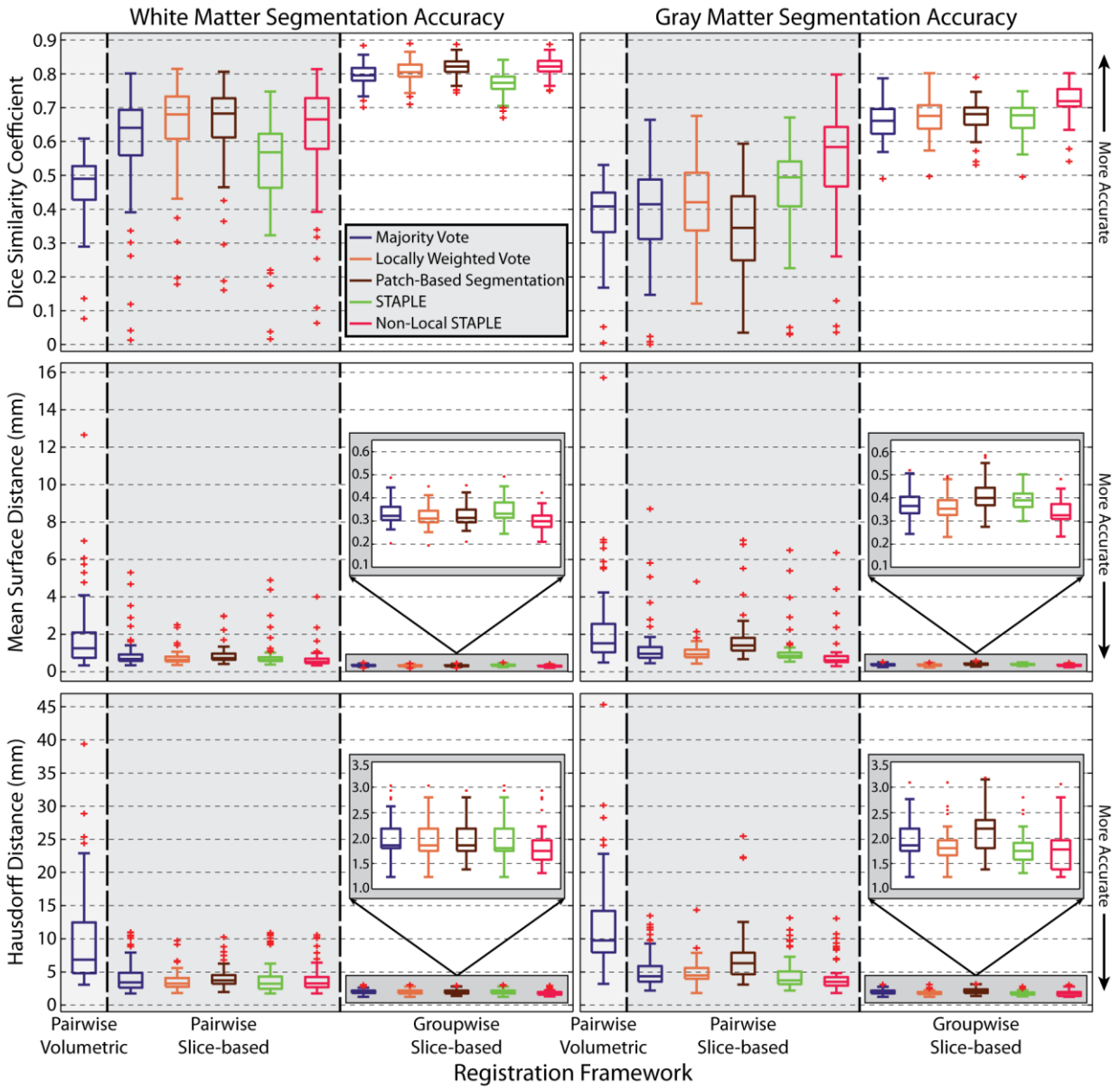


Figure VII.5. Quantitative comparison of the considered registration frameworks and label fusion approaches for the accuracy of both gray matter and white matter segmentation. For both structures, the accuracy is measured in terms of the Dice similarity coefficient, mean surface distance error, and Hausdorff distance error. The proposed groupwise slice-based registration framework provides consistent improvement across both structures and by all of the considered metrics.

and WM, respectively, and an improvement of approximately 7mm and 10mm in HDE error for the GM and WM, respectively.

As expected, the label fusion approach significantly impacts overall segmentation accuracy. However, perhaps surprisingly, no label fusion algorithm was statistically significantly better across all

metrics and anatomy considered. NLS appears to be the most consistent performer as it results in statistically significant improvement (paired t-test, $p < 0.01$) over the other considered fusion algorithms in terms of 4 of the 6 metrics – DSC and MSDE for GM, and MSDE and HDE for WM.

Both the slice-based (Figure VII.6) and volumetric (Figure VII.7) qualitative analysis corroborate the quantitative improvement provided by the proposed registration framework. In Figure VII.6, two representative slices are presented for both the proposed groupwise and the baseline pairwise slice-based registration frameworks. The first example (top of Figure VII.6) provides an example where inconsistencies in the pairwise registrations result in highly inaccurate and inconsistent segmentation results. To contrast, the corresponding groupwise results are substantially closer to the manual labels despite the lack of contrast on the corresponding target image. The second example (bottom of Figure VII.6) shows one of the best-case scenarios for the slice-based pairwise registration framework. Nevertheless, the proposed groupwise framework is consistently more accurate in its ability to maintain

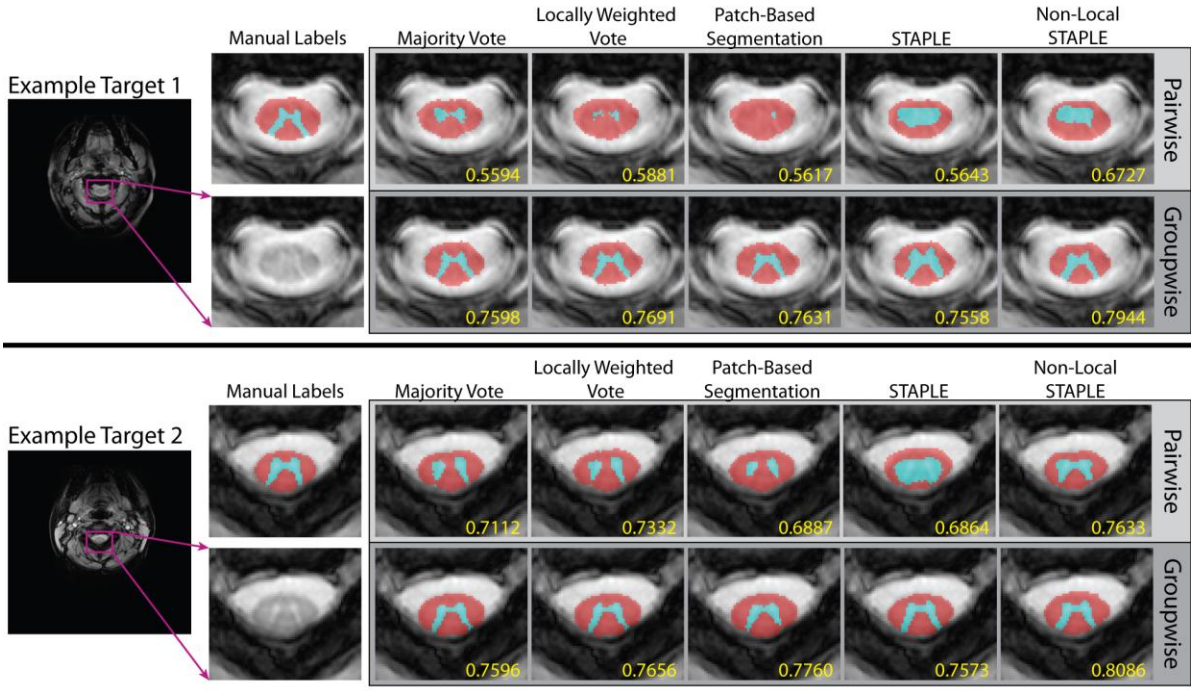


Figure VII.6. Slice-based qualitative comparison of a pairwise slice-based registration framework and the proposed groupwise slice-based registration framework. For both examples, the proposed framework provides significantly more accurate segmentations and is able to maintain the complex structure of the GM horn.

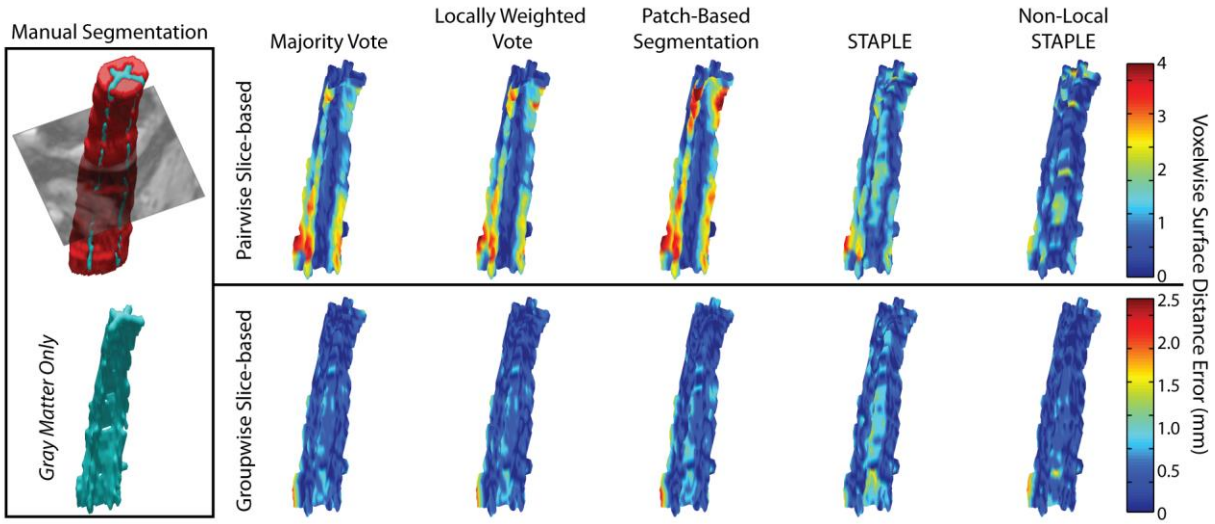


Figure VII.7. Volumetric qualitative comparison of the accuracy of the segmented gray matter for the pairwise slice-based framework, and the proposed groupwise slice-based framework. The proposed registration framework consistently estimates the complex shape of the GM horn more accurately than its pairwise counterpart. Note the different axes for the two different registration frameworks.

the shape of the GM structure and, thus, results in substantially more accurate segmentation estimates. Likewise, the representative volumetric example (Figure VII.7) illustrates consistent improvement in terms of voxelwise surface distance error for the estimation of the GM horn. In particular, it is evident that all of the fusion algorithms are capable of consistently estimating the complex shape of the GM structure throughout the volume, given the proposed groupwise registration framework. Note the different scale on the axes for the two registration frameworks, which was used to provide finer detail on the voxelwise accuracy of the GM segmentations.

3.5.2. Sensitivity to Model Parameters

The accuracy of the proposed registration framework is not particularly sensitive to either of the considered model parameters (Figure VII.8). In terms of sensitivity to the fraction of explained model variance (Figure VII.8A), it is clear that, in general, increasing the number of modes of variation provides valuable benefits in terms of segmentation accuracy. Thus, it can be inferred that the complex appearance of the spinal cord requires the use of a significant fraction of the explained variance in order to accurately approximate the relationships between the projected target and the groupwise representation of the

atlases. However, utilizing too many modes of variation (e.g., $\kappa = 0.9999$) results in sub-optimal performance as these additional modes are modeling the noise in the atlases, as opposed to structurally-meaningful spinal cord variability.

In terms of the sensitivity to the model weighting parameter, τ, it is evident that the estimated parameter value (τ_{est} in eq. 16) properly scales the geodesic relationships between the projected model image and the projected atlas information (Figure VII.8B). As a result, when $\tau - \tau_{est} > 0$, the approach is more selective, and inaccurately discards the impact of atlases that are relatively close in the low-

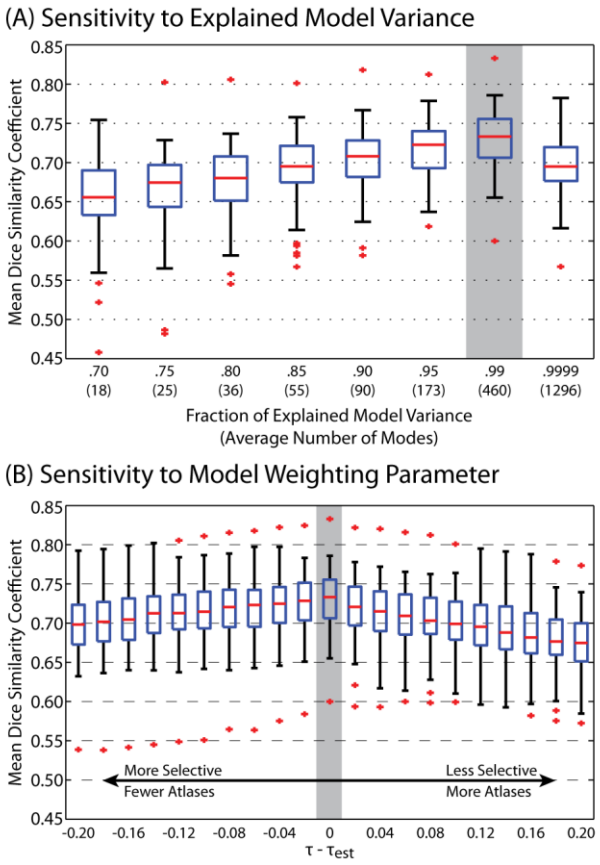


Figure VII.8. Quantitative analysis of the sensitivity of the proposed registration framework to the free model parameters. For the fraction of explained variance, (A), the inclusion of a significant portion of the modes of variation provides valuable benefits in terms of segmentation accuracy (i.e., up to $\kappa = 0.99$). However, inclusion of too many modes (i.e., $\kappa = 0.9999$) results in sub-optimal performance. For the model weighting parameter, (B), the estimated parameter value inferred from the model appears to be the near-optimal parameter value across the considered targets. For both parameters, the gray bar indicates the value used in the other presented experiments. Note, for (A) and (B) the accuracy measures are the result of a majority vote so that the effect of the parameter is not obfuscated by more sophisticated fusion algorithms.

dimensional model space. Alternatively, when $\tau - \tau_{est} < 0$, the impact of unrepresentative atlases is disproportionately large in the registration and segmentation results, resulting in sub-optimal segmentation accuracy. Interestingly, inclusion of fewer atlases (i.e., $\tau - \tau_{est} > 0$) is more detrimental to segmentation accuracy than the inclusion of extra atlases (i.e., $\tau - \tau_{est} < 0$). As a result, it can be inferred that the inclusion of extra atlases in the fusion process provides valuable improvement in segmentation accuracy – a strong indicator for the complexity and variability of the human cervical spinal cord.

4. Discussion

We propose a groupwise slice-based registration framework through the construction of an appearance model representation of the cervical spinal cord (Figures VII.2-VII.4). The proposed framework provides a powerful mechanism for: (1) modeling the variability exhibited within the cervical spinal cord population (Figure VII.2), (2) naturally and rapidly registering a target slice to the modeled spinal cord population (Figure VII.4), and (3) selecting a collection of geodesically appropriate atlases for segmenting the target image (Figures VII.3 and VII.4). While none of the individual components of the proposed framework are fully unique to this work, together, these techniques enable the first fully-automated framework for robust and accurate segmentation of the cervical spinal cord's internal structure. We have demonstrated superior performance over typical pairwise (volumetric and slice-based) multi-atlas registration frameworks. Quantitatively, we demonstrate significant segmentation accuracy improvements for both GM and WM segmentation across five different fusion approaches and three different accuracy metrics (Figure VII.5). Additionally, qualitative segmentation assessment supports the quantitative improvement in both slice-based (Figure VII.6) and volumetric (Figure VII.7) representations.

The sensitivity of the proposed model-based groupwise representation of the spinal cord was assessed with respect to the primary model parameters. The results of this sensitivity analysis (Figure VII.8) demonstrate that the proposed approach is not particularly sensitive to (1) the fraction of explained

variance (Figure VII.8A) or (2) the weighting parameter for determining the optimal collection of representative atlases (Figure VII.8B). The results of this analysis provide several important insights into the robustness of the proposed framework. First, modeling the complex inter-subject variability in the spinal cord requires a significant fraction (i.e., $\kappa = 0.99$) of the explained model variance to be utilized. However, not surprisingly, the inclusion of nearly all of the modes of variation (i.e., $\kappa = 0.9999$) has a detrimental impact on registration/segmentation accuracy. Secondly, the inclusion of extra atlases by means of a larger model weighting parameter, τ, is less detrimental than the inclusion of fewer atlases, which strongly implies the utility of selecting appropriate and sufficient atlases when modeling the complexity and variability of the human cervical spinal cord.

Although label fusion is certainly not the focus of this manuscript, it is worth noting the fact that none of the fusion algorithms used in this manuscript reported superior performance across all of the considered metrics (Figure VII.5). While NLS was the most consistent performer, resulting in statistically significant improvement in 4 of the 6 considered metrics, the lack of a consistently superior fusion algorithm is problematic. Thus, investigation into label fusion optimality (e.g., through re-formulating rater models [50, 52, 54, 55] or corrective learning [69, 70]) remains an open problem and certainly warrants continued investigation.

Despite the promise of the proposed framework, there are several areas for future investigation that could provide increased applicability to new problem spaces. First, all of the registrations performed in this manuscript were simple, 2-D (three degree-of-freedom) rigid transformations. Additional degrees of freedom (e.g., scale, skew) and/or deformable registration techniques would enable a more compact representation (i.e., fewer modes of variation) of the cervical spinal cord variability. However, this more compact representation would come at the cost of more complex parameter optimization, and, thus, increased likelihood in converging to an undesired local minimum. In our experiments, the three degree-of-freedom approach was found to (1) succinctly model observed spinal cord variability, and (2) quickly and robustly find inter-subject correspondence. Second, the proposed framework was performed entirely on the 2-D cross-section of the spinal cord, without regard for introducing or enforcing 3-D consistency.

Consistency throughout the image volume could theoretically be maintained through the use of (1) Markov Random Fields [18, 198], or (2) constraints on the slice-based rigid transformations [199]. Due to the consistent performance provided by the proposed framework and concerns with the degree of anterior-posterior image distortion, we did not feel that this was necessary at this time. Lastly, and probably most obviously, further investigation into modeling complex 3-D structures (as opposed to the 2-D approach presented here) would increase the applicability of the proposed groupwise approach.

Finally, given the known anatomical context of the spinal canal and the vast research that has gone into the optimality and design of non-rigid registration algorithms, there is no doubt that 3-D volumetric registrations could be more successful than the results presented in this manuscript (Figures VII.1 and VII.5). However, one of the primary motivations for this work was to demonstrate that typical approaches, which are often highly successful on oft-studied structures (e.g., the brain), are problematic when applied to new, highly difficult structures (i.e., structures exhibiting large imaging and anatomical variability). As a result, reasonable segmentation of the spinal cord's internal structures through 3-D deformable registrations often require (1) *a priori* structural information, (2) a highly-tuned application-specific registration framework – e.g., [200], or (3) a multi-contrast cost function (e.g., using T1- and T2*-weighted images). Additionally, and potentially most importantly, pairwise 3-D non-rigid registration algorithms can often take upwards of an hour to perform each individual registration on a modern CPU. For the LOOCV presented above, estimating a complete 3-D segmentation took almost three days of CPU time per target. Given offline construction of the appearance model, the framework presented in this manuscript took approximately one minute per target slice, resulting in a complete 3-D segmentation in approximately 30 minutes – less than the time it would take to perform a single atlas-target non-rigid registration.

CHAPTER VIII

GEODESIC LEARNER FUSION

1. Overview

Multi-atlas segmentation is a powerful generalize-from-example framework for image segmentation [9, 26]. In multi-atlas segmentation, a set of labeled atlases are non-rigidly registered to a target image [36, 101] and the resulting label conflicts are resolved using label fusion [61, 78]. Due to the robustness and lack of anatomical assumptions, multi-atlas segmentation has grown tremendously over the past decade. Unfortunately, this robustness comes at the cost of computational complexity as typical multi-atlas approaches rely on an expensive pairwise registration framework. While these independent registrations play a critical role in overcoming the deficiencies in the individual atlas observations, each registration can take on the order of hours to converge to an acceptable local correspondence.

We propose a whole-brain (133 label) multi-atlas segmentation framework using a big data paradigm. Building on seminal works in machine learning (e.g., AdaBoost [114] and Principal Component Analysis – PCA), we use a learning-based approach to emulate the accuracy of a premier multi-atlas segmentation framework while dramatically lessening the computational burden. Given a large collection of training data which was pre-processed using a state-of-the-art multi-atlas segmentation procedure, we: (1) construct a low-dimensional representation of our training data for computing neighborhood relationships and (2) optimize an AdaBoost classifier for each training image that maps a weak segmentation estimate (e.g., a majority vote of the geodesic neighbors) to the expensive, yet highly accurate, multi-atlas segmentation estimate. Thus, when a new target image needs to be segmented we simply need to (1) project the image into the low-dimensional space, (2) construct a weak initial

segmentation, and (3) fuse the geodesically appropriate learners from the training phase. We refer to the algorithm as geodesic learner fusion (GLF) – Figure VIII.1.

2. Data and Pre-Processing

Herein, the complete data aggregates 7 unique datasets (4 of which are publicly available) covering a wide range of demographics, ages, and neurological states (Table VIII.1). In total, a set of 3,505 subjects were scanned resulting in a total of 3,886 T1-weighted MR whole-brain volumes. For validation, the data was separated into three groups: *training*, *testing*, and *reproducibility*. First, the MMMRR dataset was used in its entirety as the reproducibility set as it consists of 21 subjects identically scanned twice. The remaining datasets were split 90%/10% into the training/testing cohorts. Note, all intra-subject scans were placed accordingly in the same training/testing group.

For all 3,886 images, a state-of-the-art multi-atlas segmentation was performed. The original atlases are a set of 45 MPRAGE images (from unique subjects) as part of the OASIS dataset [145]. All atlases were labeled with 133 labels (BrainCOLOR protocol [154]). For consistency, all images were affinely registered [101] to the MNI305 atlas [86]. For each image, the 15 closest atlases were selected

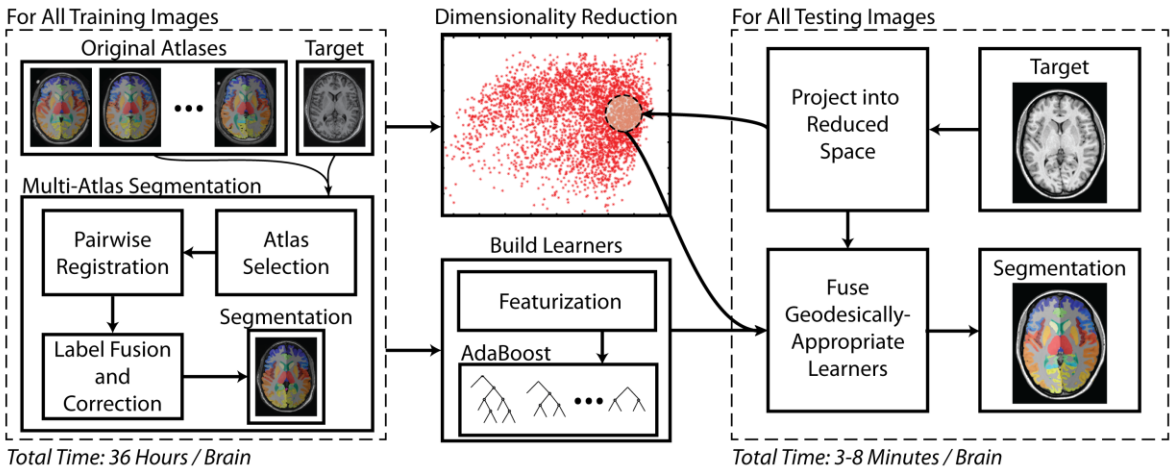


Figure VIII.1. Flowchart demonstrating the geodesic learner fusion (GLF) framework. A large collection of training images are processed offline using a typical multi-atlas segmentation pipeline. The dimensionality of the training images is then reduced, and learners are constructed to map a weak initial estimate to the multi-atlas segmentation. Finally, for a new testing image, the image needs to be projected into the low-dimensional space and the geodesically appropriate learners can be fused to efficiently and accurately estimate the final segmentation.

Table VIII.1. Data summary. Each value is represents: number of subjects (number of images)

	Training	Testing	Reproducibility
1000 Functional Connectome (fcon_1000)[a]	1055 (1055)	117 (117)	
Baltimore Longitudinal Study on Aging (BLSA)	578 (883)	64 (94)	
Information eXtraction from Images (IXI)[b]	523 (523)	58 (58)	
Deep Brain Stimulation (DBS)	493 (493)	54 (54)	
Open Access Series on Imaging Studies (OASIS)[c]	375 (392)	41 (44)	
Tennessee Twins Study (TTS)	113 (118)	13 (13)	
Multi-Modal MRI Reproducibility Resource (MMMRR)[d]			21 (42)
Total:	**3137 (3464)**	**347 (380)**	**21(42)**

a: https://www.nitrc.org/projects/fcon_1000/
b: http://www.oasis-brains.org/
c: http://biomedic.doc.ic.ac.uk/brain-development/
d: https://www.nitrc.org/projects/multimodal

(using a naïve PCA projection), pairwise registered [36, 101], fused [52, 78], and corrected through implicit error modeling [70]. On average, this process took 36 hours on a modern computer.

Finally, for all 3464 training images, a low-dimensional representation was computed using PCA. Briefly, all images were down-sampled to 2mm isotropic and the principal components were computed over the brain region defined by the multi-atlas segmentation estimates. Geodesic distances are computed using the projection weights onto the first 15 modes of variation (representing 15.33% of the total variation). The results of the pre-processing framework are summarized in Figure VIII.2.

3. Geodesic Learner Fusion Theory

The theory presented below builds on the foundation for learning-based error correction presented in [70]. For training image j, we assume that we are given (1) the target image, $\boldsymbol{I}_j \in \mathbb{R}^N$, (2) the initial

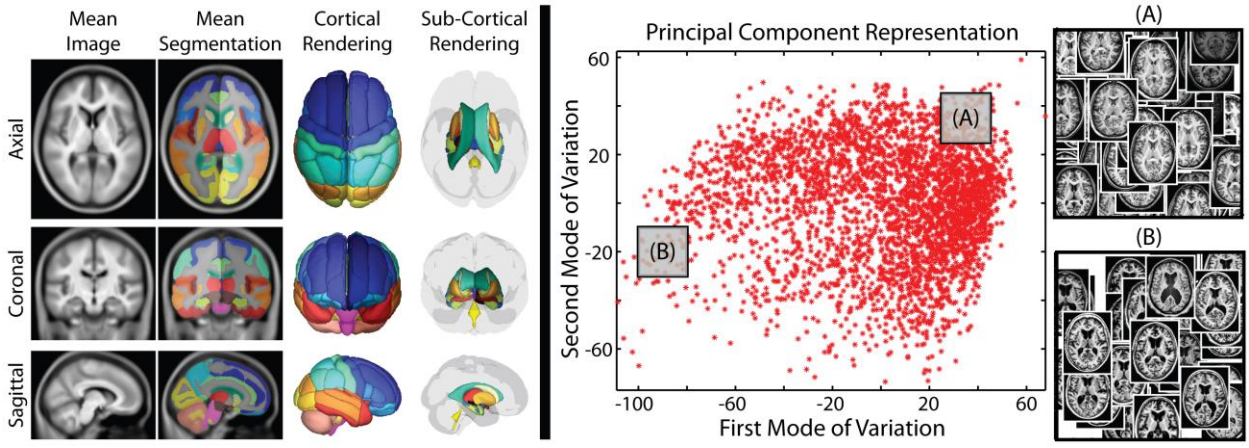


Figure VIII.2. Summary of the training data processed through multi-atlas segmentation and their corresponding representation in the estimated low-dimensional space. The inlays in (A) and (B) illustrate that the geodesic distance metric leads to clustering of similar anatomical features.

weak segmentation, $\boldsymbol{\Psi}_j \in \boldsymbol{L}^N$, and (3) the multi-atlas segmentation, $\boldsymbol{\Omega}_j \in \boldsymbol{L}^N$, where N is the total number of voxels, and $\boldsymbol{L}$ is the set of possible labels (herein, $|\boldsymbol{L}| = 133$). As in [70], the AdaBoost training procedure is computed for all of the labels independently. For each label, let $\boldsymbol{B}_l$, such that $l \in \boldsymbol{L}$, be the collection of voxels for which any of the training images observe label l.

For the classifier, let the feature matrix be defined as $\boldsymbol{X}^l \in \mathbb{R}^{M \times F}$, such that each element, X^l_{mf}, is the feature value for feature f at sample m and label l, where F is the number of features, and $M \leq |\boldsymbol{B}_l|$ is the number of samples (or voxels). For simplicity, we define the features at each sample the same way as [70]. Briefly, these consist of the voxel coordinates, the observed labels (i.e., all $\Psi_{ji}\ s.t.\ i \in \boldsymbol{R}_m$), the target intensities (i.e., all $\mathrm{I}_{ji}\ s.t.\ i \in \boldsymbol{R}_m$), and the corresponding spatial correlations – where $\boldsymbol{R}_m$ is the collection of voxels within the feature window defined for sample m (herein, a 5mm isotropic window centered at the current sample). This feature collection strategy results in a total number of features of $F = 1009$. Finally, we define the class vector as, $\boldsymbol{Y}^l \in \{-1, 1\}^M$, where each element $Y^l_m = 1$ if $\Omega_{jm} = l$, and $Y^l_m = -1$ otherwise.

For the AdaBoost training, let $\boldsymbol{D}^{(t)}_{jl} \in \mathbb{R}^M$, be the distribution of relative weights for all samples at iteration $t \leq T$ (where $D^{(0)}_{jlm} = \frac{1}{M}$ initially). The goal of the training process at iteration t is to optimize the weak learner, h_{jlt}, where $h_{jlt}[X^l_m] \in \{-1,1\}$

$$h_{jlt} = \arg\max_{h_{jlt}} \left| 0.5 - \sum_m D^{(t)}_{jlm} \left(1 - \delta\left(h_{jlt}[X^l_m], Y^l_m\right)\right) \right| \tag{8.1}$$

where, $\delta(\cdot,\cdot)$ is the Kronecker delta function. Note, herein, the weak learner in (1) is a decision tree and optimization of this learner is addressed later in the manuscript. Next, the weight associated with the current iteration, $\alpha_{jlt} \in \mathbb{R}$, is defined as

$$\alpha_{jlt} = \frac{1}{2} \ln \frac{1 - \sum_m D^{(t)}_{jlm} \left(1 - \delta\left(h_{jlt}[X^l_m], Y^l_m\right)\right)}{\sum_m D^{(t)}_{jlm} \left(1 - \delta\left(h_{jlt}[X^l_m], Y^l_m\right)\right)} \tag{8.2}$$

and the sample weight can be updated with

$$D_{jlm}^{(t+1)} = \frac{1}{Z}\exp\left(\alpha_{jlt}\delta\left(h_{jlt}[X_m^l], Y_m^l\right)\right) \tag{8.3}$$

where Z is a partition function ensuring that $\sum_m D_{jlm}^{(t+1)} = 1$. This process is then iterated until we have reached the desired number of iterations, T (herein, $T = 50$).

Once the training process has been performed on all training images, we can then approximate the desired multi-atlas segmentation through fusing the trained AdaBoost learners associated with the corresponding geodesically selected training images. If we let $\boldsymbol{J}$ be the set of selected training images, and $\boldsymbol{\Omega}^* \in \boldsymbol{L}^N$ be the approximated multi-atlas segmentation, then Ω_i^* (i.e., the estimated label at voxel i) is computed as

$$\Omega_i^* = \arg\max_{l \in \boldsymbol{L}} \sum_{j \in \boldsymbol{J}} \sum_t \alpha_{jlt} h_{jlt}[X_i^l] \tag{8.4}$$

where the feature matrix, $\boldsymbol{X}$, is defined in exactly the same way for the testing image as it was previously defined for the training images.

4. Methods and Results

Throughout, all segmentation comparisons are assessed with the mean Dice Similarity Coefficient

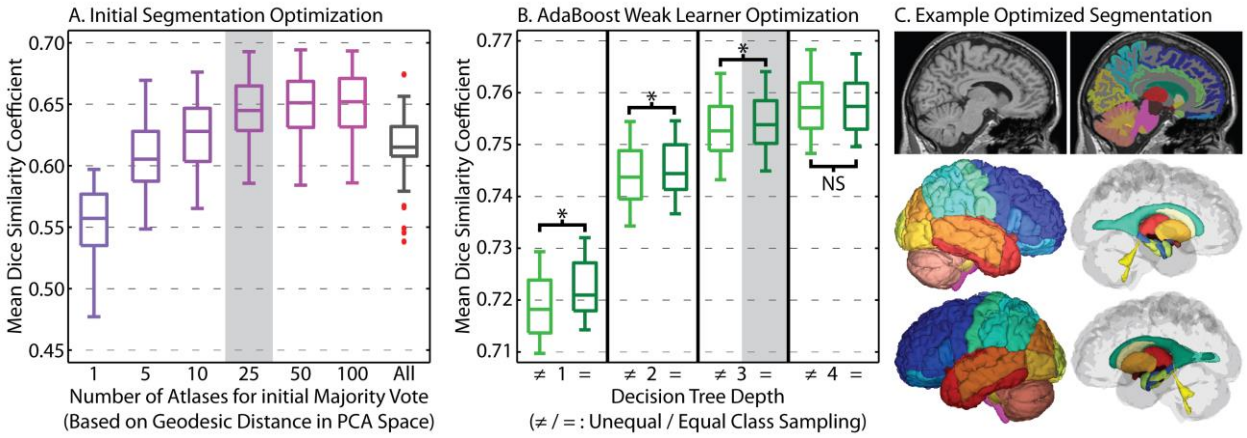


Figure VIII.3. Parameter optimization and sensitivity for the number of atlases fused for the initial majority vote (A), and the type of weak learner used for the AdaBoost classifiers (B). A representative segmentation using the optimized parameters can be seen in (C). Note, on (B), "*" indicates statistically significant difference, and "NS" indicates no significant difference.

(DSC) [140] across the 132 non-background labels, and all claims of statistical significance are made using a Wilcoxon signed rank test ($p < 0.01$) [153].

4.1. Parameter Optimization and Sensitivity

First, we optimize: (1) the number of geodesically selected atlases for the initial weak segmentation (via a majority vote), and (2) the weak learner used in the AdaBoost classifier. For optimization, the desired parameters were swept across an appropriate range for a random subset of 50 training images. The results can be seen in Figure VIII.3. For the initial majority vote accuracy (Figure VIII.3A), using too few (e.g., 5) or too many (e.g., all available training data) results in sub-optimal accuracy. Additionally, there is marginal return when increasing the number of selected atlases beyond 25. Thus, as computation time is of primary concern, the geodesically closest 25 training images were used for all subsequent analysis. For the AdaBoost weak learner optimization (Figure VIII.3B), we consider decision trees with depths ranging from 1 (i.e., a "decision stump") to 4. Additionally, we consider two sampling methods, *unequal* and *equal*. For unequal sampling, all available samples were

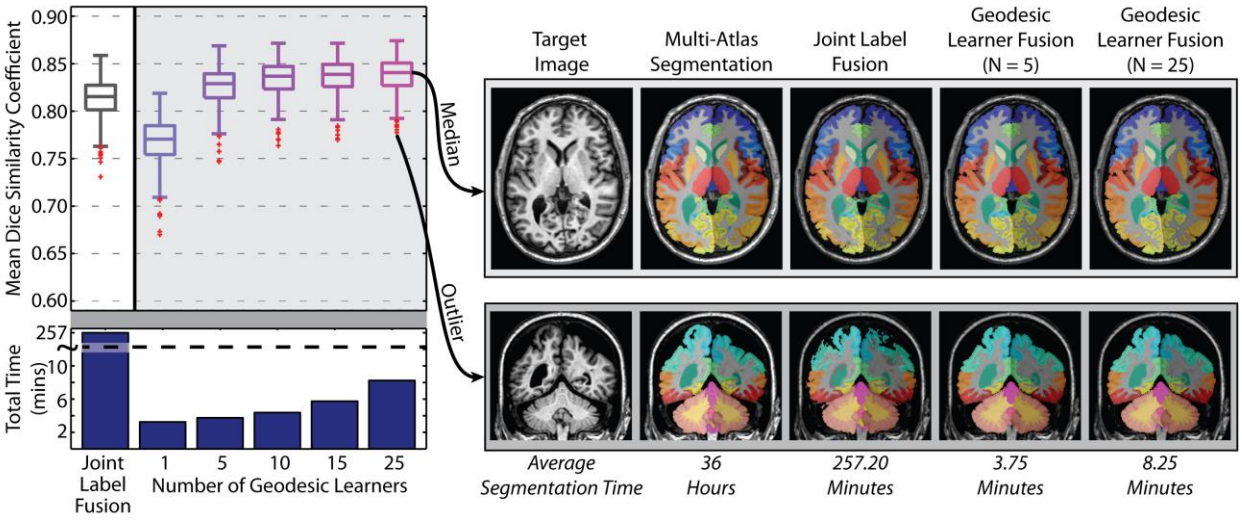


Figure VIII.4. Mean accuracy assessment for the defined testing data using the multi-atlas segmentation estimate as a "silver standard". The results demonstrate (1) the GLF framework provides a dramatic decrease in total segmentation time, (2) increasing the number of fused learners has valuable benefits in terms of segmentation accuracy, and (3) when fusing more than 5 geodesic learners the GLF framework provides substantial and significant accuracy benefits over the joint label fusion baseline.

used for each label, regardless of the resulting class imbalance. For equal sampling, a random subset of the possible samples was selected to enforce class balance. Here, it is evident that (1) increasing the decision tree depth improves training accuracy, and (2) equal class sampling provides a marginal, yet significant, improvement in segmentation accuracy. Given the marginal return and dramatic runtime increase of a depth 4 decision tree, a depth 3 decision tree with equal class sampling was used for all subsequent experiments.

4.2. Testing Data Accuracy and Assessment

Next, we quantify our ability to replicate the expensive multi-atlas segmentation result using the GLF framework. Using the multi-atlas segmentation estimate on our testing data (380 images) as a "silver standard" we applied the GLF framework with varying numbers of geodesic learners (from 1 to 25). As a benchmark, we consider fusing the 25 geodesically nearest training images using the premier joint label fusion (JLF) algorithm [61]. The results of this experiment across the 380 testing images (Figure VIII.4) demonstrate: (1) increasing the number of geodesic learners results in an improved ability to replicate the multi-atlas segmentation result, (2) using at least 5 learners results in significant and substantial improvement over the JLF benchmark, and (3) increasing the number of learners from 1 to 25 increases the total segmentation time from approximately 2 minutes to approximately 8 minutes – which remains a speedup of ≈ 31x over the JLF benchmark and ≈ 262x over the multi-atlas framework. The qualitative results support the quantitative accuracy analysis for both the worst and median cases from the testing set.

4.3. Reproducibility Data Accuracy and Assessment

Lastly, we assess the reproducibility of the GLF framework using the MMMRR dataset (see Table VIII.1). Within this dataset, all 21 subjects were scanned twice with exactly the same scanning parameters. The intra-subject reproducibility was assessed by comparing the mean DSC for: (1) the GLF result vs. the corresponding multi-atlas result, (2) the intra-subject multi-atlas estimates, and (3) the intra-subject GLF framework estimates. The results (Figure VIII.5) demonstrate: (1) the GLF similarity to the

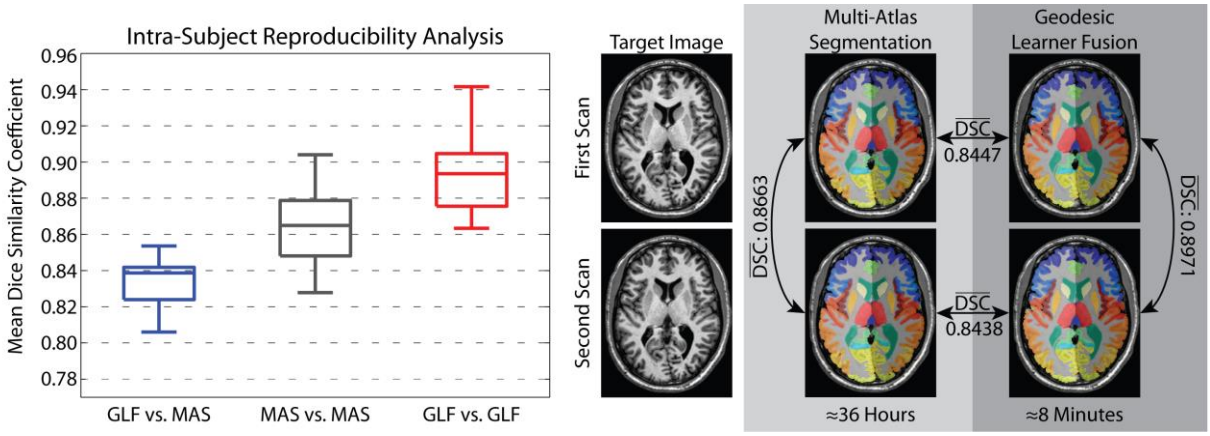


Figure VIII.5. Reproducibility analysis on the MMMRR dataset. Note, (1) the GLF similarity to the multi-atlas segmentation result approaches the intra-subject reproducibility for multi-atlas segmentation, and (2) GLF is significantly more reproducible than multi-atlas segmentation on this dataset.

multi-atlas segmentation result approaches the intra-subject reproducibility for multi-atlas segmentation, and (2) GLF is significantly more reproducible than multi-atlas segmentation with a mean intra-subject DSC improvement of 0.0288.

5. Discussion

We present geodesic learner fusion (GLF), a framework for replicating the robust and accurate multi-atlas segmentation model, while dramatically lessening the computational burden. Using a training set of 3464 images, we estimate a low-dimensional representation of brain anatomy for selecting geodesically appropriate example images, and build AdaBoost learners that map weak initial segmentations to the more accurate multi-atlas segmentation result. By completely bypassing the deformable atlas-target registrations, the GLF framework, cuts the runtime on a modern computer from 36 hours down to 3-8 minutes – a speedup that could be further enhanced through GPU-based optimization. Specifically, we: (1) describe a technique for optimizing the initial segmentation and the AdaBoost learning parameters (Figure VIII.3), (2) quantify the ability to replicate the multi-atlas result with mean DSC of approximately 0.85 on a testing set of 380 images (Figure VIII.4), and (3) demonstrate accuracies that are approaching the intra-subject multi-atlas reproducibility on a separate reproducibility dataset, and show significant increases in GLF reproducibility (Figure VIII.5).

In the interest of brevity, all of our comparisons have been against the standard pairwise registration framework for multi-atlas segmentation, and have not included the more recent advancements in groupwise registration (e.g., [188]). The primary reason for not directly including this comparison is: (1) groupwise registration is still a very active area of continuing research, and (2) the GLF framework is, in its essence, a machine learning perspective on the groupwise registration model.

In the end, while the GLF framework shows great promise for rapid and accurate multi-atlas segmentation, there are certainly areas for which further investigation is warranted. Namely, first, we used a naïve PCA projection to model the geodesic relationships between the training images. More recent advancements in the manifold learning literature (e.g., [201]) present fascinating opportunities for more accurately modeling these relationships. Second, while highly successful, we do not claim any optimality of our AdaBoost-based learners. Investigation into alternative classification techniques (e.g., [76]) could provide valuable improvements in segmentation modeling without dramatically altering the GLF framework.

CHAPTER IX

CONCLUSIONS AND FUTURE WORK

1. Summary

The ability to generalize information from examples has been the driving force behind decades of statistical modeling and machine learning research. Building on this fundamental concept, this dissertation addresses the ability to generalize structural context, or segmentations, from medical images using labeled examples (i.e., atlases). Specifically, this research focuses on the problem of multi-atlas segmentation, in which image correspondences between a set of atlases and the target-of-interest are discovered and the underlying target segmentation is estimated using statistical fusion – a supervised learning approach for resolving label conflicts. Using this general framework, several theoretical advancements to the statistical fusion model are presented (**Chapters II-V**), and the results of these contributions are highlighted on clinically and scientifically relevant applications (**Chapters VI-VIII**).

2. Theoretical Advancements to Statistical Fusion

2.1. Summary

We present theoretical reformulations to the statistical fusion framework to more accurately characterize rater (or atlas) performance (**Part 1**). Specifically, these advancements provide methods for: (1) estimating task difficulty (**Chapter II**), (2) formulating spatially varying performance (**Chapter III**), (3) accounting for registration uncertainty and imperfect correspondence (**Chapter IV**), and (4) estimating hierarchically consistent models of performance (**Chapter V**). Together, these theoretical advancements provide powerful mechanisms for more accurately understanding and estimating rater-driven models and, thus, more accurately estimating the desired target segmentations.

2.2. Main Contributions

1. We present COLLATE, an algorithm for fusing a collection of rater label observations to estimate the consensus level, labeler accuracy and truth labels. Like its predecessors, COLLATE takes a collection of input observations from a group of raters (human or otherwise) and simultaneously estimates the truth labels and the rater performance parameters ("labeler accuracy"). However, COLLATE also estimates the consensus level of each voxel, which can be viewed as an inherent property of each voxel that determines the likelihood that a given rater would be confused about the label associated with a given voxel.
2. We present Spatial STAPLE — a new algorithm for statistically fusing rater label information using a spatially varying model of rater behavior. Spatial STAPLE: (i) provides significant improvement over the premier label fusion techniques, (ii) more accurately reflects the way in which raters and atlases make mistakes than traditional global performance metrics, and (iii) provides a unified framework that can be used for the gamut of label fusion applications (i.e. the fusion of human raters, multi-atlas applications and the fusion of multiple algorithms).
3. We present Non-Local STAPLE, a statistical fusion algorithm for multi-atlas segmentation. Through a reformulation from a non-local means perspective, NLS represents the first statistical fusion algorithm that (i) creates a cohesive theoretical model specifically targeting registered atlas observation behavior, and (ii) seamlessly incorporates intensity into the core of the STAPLE estimation framework. As a result, NLS largely overcomes the need for high-quality non-rigid registration and large numbers of atlases.
4. We propose a novel statistical fusion framework using a reformulated hierarchical performance model. Given an *a priori* model of the hierarchical label relationships for a given segmentation task, the proposed generative model of rater performance provides a straightforward mechanism for quantifying rater performance at each level of the hierarchy. The primary contributions of this work are we: (i) provide a theoretical advancement to the

statistical fusion framework that enables the simultaneous estimation of multiple (hierarchical) confusion matrices for each rater, (ii) show that the proposed hierarchical formulation is highly amenable to many of the state-of-the-art advancements that have been made to the statistical fusion framework, and (iii) demonstrate statistically significant improvement on both simulated and empirical data.

2.3. Future Work

While the proposed advancements have provided valuable theoretical perspectives on the statistical label fusion problem, the primary remaining question is to define a unified model in which (1) task difficulty, (2) spatially varying performance, (3) non-local correspondence models, and (4) hierarchical performance models can be presented as a single statistical fusion performance model [202]. While these advancements are fully compatible with one another, exploration into the optimal combination of these advancements remains an open problem and represents fascinating avenues of continuing research.

3. Out-of-Atlas Likelihood Estimation

3.1. Summary

The ability to detect abnormalities and anomalies in medical images plays a critical role in the detection of diseases and pathologies as well as maintaining image quality assurance. Unfortunately multi-atlas segmentation is limited to "in-atlas" applications (e.g., applications where the atlases are anatomically and structurally indicative of the target image). We propose a technique to estimate the out-of-atlas (OOA) likelihood for every voxel in the target image (**Chapter VI**). The OOA approach provides an intuitive and fully general abnormality/outlier detection framework

3.2. Main Contributions

1. The proposed OOA framework extends the multi-atlas labeling paradigm to be sensitive to abnormalities present in medical images. Previous work on the problem of abnormality detection has primarily relied on a single atlas (or template) and, as a result, has been largely dependent on highly accurate non-rigid registration. The proposed method provides a fully general framework that (1) uses multiple normal atlases to limit the inherent bias of using a single atlas and avoid the need for non-rigid registration, and (2) can be used in a large number of potential applications.
2. On an empirical experiment for detecting malignant gliomas in the human brain, the OOA algorithm demonstrates a natural model for detecting large-scale abnormalities in the human brain. The proposed framework can consistently and reliably declare voxels to be cancerous in terms of an increasing declaration threshold.
3. In a second experiment, we demonstrate the ability of the proposed algorithm to be used in a DTI quality control framework. The proposed algorithm clearly detects large-scale quality control issues and provides consistently high OOA likelihoods across the observed data.

3.3. Future Work

Despite the promise of the OOA likelihood estimation framework, there are limitations to the proposed approach. First, we use a collection of normal (non-gadolinium enhanced) T1-weighted atlases and use them to assess images that were acquired using clinical imaging protocols (e.g., differing imaging sequence). As a result, the ability to intensity normalize these images is limited and we are forced to limit ourselves to applications where the intensity profile of the desired abnormality is dramatically different than normal anatomy (e.g., malignant gliomas). The use of the proposed framework for the detection of more subtle anatomical pathologies would be inherently limited unless the atlases were constructed using the appropriate imaging characteristics.

4. Groupwise Segmentation of the Spinal Cord

4.1. Summary

The spinal cord is an essential and vulnerable component of the central nervous system which can be significantly affected by numerous neurological conditions. Differentiating and localizing pathology/degeneration of the gray matter (GM) and white matter (WM) plays a critical role in assessing the magnitude of tissue damage, therapeutic impacts and determining prognosis of these conditions. We propose the first approach for fully automated segmentation of cervical spinal cord internal structure using a groupwise slice-based multi-atlas registration framework (**Chapter VII**).

4.2. Main Contributions

1. We provide a method for (i) pre-aligning the slice-based atlas information into a common, groupwise-consistent coordinate system, (ii) constructing a model describing spinal cord variability (i.e., "eigenspines"), (iii) registering the target image slice to the model space using a simultaneous intensity- and model-driven cost function, and (iv) estimating a final segmentation by fusing the provided atlas information.
2. The proposed framework provides a natural mechanism for selecting geodesically appropriate atlases (i.e., atlas selection) and initializing the free model parameters in an informed model-specific context.
3. We have demonstrate superior performance over typical pairwise (volumetric and slice-based) multi-atlas registration frameworks. Quantitatively, we demonstrate significant segmentation accuracy improvements for both GM and WM segmentation across five different fusion approaches and three different accuracy metrics.

4.3. Future Work

There are several areas for future investigation that could provide increased applicability to new problem spaces. First, all of the registrations performed in this manuscript were simple, 2-D (three

degree-of-freedom) rigid transformations. Additional degrees of freedom (e.g., scale, skew) and/or deformable registration techniques would enable a more compact representation (i.e., fewer modes of variation) of the cervical spinal cord variability. However, this more compact representation would come at the cost of more complex parameter optimization, and, thus, increased likelihood in converging to an undesired local minimum. In our experiments, the three degree-of-freedom approach was found to (1) succinctly model observed spinal cord variability, and (2) quickly and robustly find inter-subject correspondence. Second, the proposed framework was performed entirely on the 2-D cross-section of the spinal cord, without regard for introducing or enforcing 3-D consistency. Due to the consistent performance provided by the proposed framework and concerns with the degree of anterior-posterior image distortion, we did not feel that this was necessary at this time. Lastly, and probably most obviously, further investigation into modeling complex 3-D structures (as opposed to the 2-D approach presented here) would increase the applicability of the proposed groupwise approach.

5. Geodesic Learner Fusion

5.1. Summary

We propose geodesic learner fusion (GLF), a framework for rapidly and accurately replicating the highly accurate, yet computationally expensive, multi-atlas segmentation framework based on fusing geodesically appropriate learners (**Chapter VIII**). In the largest whole-brain multi-atlas study ever reported, multi-atlas segmentations are estimated for a training set of 3,464 MR brain images. Using these mulit-atlas estimates we (1) estimate a low-dimensional representation for selecting geodesically appropriate example images, and (2) build AdaBoost learners that map a weak initial segmentation to the multi-atlas segmentation result. Thus, to segment a new target image we simply project the image into the low-dimensional space, construct a weak initial segmentation, and fuse the trained, geodesically appropriate, learners.

5.2. Main Contributions

1. We describe a technique for optimizing the weak initial segmentation and the AdaBoost learning parameters. Resulting in a framework that can be optimally trained in the largest whole-brain multi-atlas study ever reported (3,886 total whole brain MR images).
2. We quantify the ability to replicate the multi-atlas result with mean DSC of approximately 0.85 on a testing set of 380 images.
3. We demonstrate accuracies that are approaching the intra-subject multi-atlas reproducibility on a separate reproducibility dataset, and show significant increases in GLF reproducibility when compared to a state-of-the-art multi-atlas segmentation framework.
4. By completely bypassing the need for deformable atlas-target registrations, the GLF framework, cuts the runtime on a modern computer from 36 hours down to 3-8 minutes – a 262x speedup.

5.3. Future Work

In the end, while the GLF framework shows great promise for rapid and accurate multi-atlas segmentation, there are certainly areas for which further investigation is warranted. Namely, first, we used a naïve PCA projection to model the geodesic relationships between the training images. More recent advancements in the manifold learning literature present fascinating opportunities for more accurately modeling these relationships. Second, while highly successful, we do not claim any optimality of our AdaBoost-based learners. Investigation into alternative classification techniques could provide valuable improvements in segmentation modeling without dramatically altering the GLF framework.

6. Concluding Remarks

In early 2010, multi-atlas segmentation was growing in popularity because of its ability robustly label difficult anatomy, but was typically considered to be less accurate than parametric segmentation models and relegated to applications for which automated algorithms had not previously been developed.

In particular, the label fusion component of multi-atlas segmentation was limited to voting and weighting voting based approaches which (1) required a large number of representative atlases, and (2) resulted in sub-optimal accuracy when compared to state-of-the-art segmentation approaches. Since that time, however, the label fusion field has become a prominent area of continuing research, and optimal statistical and probabilistic models of atlas performance and atlas voting have been derived. As a result, the gap between highly optimized parametric segmentation algorithms and the generalized multi-atlas segmentation framework has dramatically closed. This dissertation has played a pivotal role in advancing the statistical label fusion theory and, thus, advancing the applicability of multi-atlas segmentation to new, previously ignored problem spaces.

Moving forward, the momentum of multi-atlas segmentation seems to be heading towards applications for which discovering dense and accurate correspondence is increasingly difficult (e.g., abdomen, cardiac). Moreover, we see an increased emphasis on (1) accurate and applicable label fusion models for atlas selection and atlas performance, and (2) building shape/appearance models of anatomical variability to avoid confounding factors in regions of low contrast (and/or high noise). Given this momentum, we believe the benefits of this dissertation have not fully been discovered and should remain relevant as the field of multi-atlas segmentation extends outside of the cranial vault and into new and more challenging anatomies.

APPENDIX A

PUBLICATIONS

1. Refereed Journal Articles

1. **Andrew J. Asman**, Bennett A. Landman, *"Hierarchical Performance Estimation in the Statistical Label Fusion Framework"* Medical Image Analysis. Submitted February 2014
2. **Andrew J. Asman**, Frederick W. Bryan, Seth A. Smith, Daniel S. Reich, Bennett A. Landman, *"Groupwise Multi-Atlas Segmentation of the Spinal Cord's Internal Structure"* Medical Image Analysis. January 2014
3. Frederick W. Bryan, Zhoubing Xu, **Andrew J. Asman**, Wade M. Allen, Daniel S. Reich, and Bennett A. Landman. *"Self-Assessed Performance Improves Statistical Fusion of Image Labels."* Medical Physics. January 2014
4. Carolyn B. Lauzon, **Andrew J. Asman**, Michael L. Esparza, Scott S. Burns, Qiuyun Fan, Yurui Gao, Adam W. Anderson, Nicole Davis, Laurie E. Cutting, Bennett A. Landman. *"Simultaneous Analysis and Quality Assurance for Diffusion Tensor Imaging"* PLoS ONE. April 2013
5. **Andrew J. Asman**, Lola B. Chambless, Reid C. Thompson, Bennett A. Landman, *"Out-of-Atlas Likelihood Estimation using Multi-Atlas Segmentation"* Medical Physics. March 2013
6. **Andrew J. Asman** and Bennett A. Landman. *"Non-Local Statistical Label Fusion for Multi-Atlas Segmentation",* Medical Image Analysis. December 2012
7. Zhoubing Xu, **Andrew J. Asman**, Eesha Singh, Lola Chambless, Reid Thompson, Bennett A. Landman, *"Segmentation of Malignant Gliomas through Remote Collaboration and Statistical Fusion"* Medical Physics. August 2012

8. **Andrew J. Asman** and Bennett A. Landman, *"Formulating Spatially Varying Performance in the Statistical Fusion Framework"* IEEE Transactions on Medical Imaging. June 2012
9. Bennett A. Landman, **Andrew J. Asman**, Andrew G. Scoggins, John A. Bogovic, Joshua A. Stein, and Jerry L. Prince. *"Foibles, follies, and fusion: Web-Based Collaboration for Medical Image Labeling"* Neuroimage. August 2011
10. Bennett A. Landman, **Andrew J. Asman**, Andrew G. Scoggins, John A. Bogovic, Fangxu Xing, and Jerry L. Prince. *"Robust Statistical Fusion of Image Labels"* IEEE Transactions on Medical Imaging. October 2011
11. **Andrew J. Asman** and Bennett A. Landman. *"Robust Statistical Label Fusion through Consensus Level, Labeler Accuracy and Truth Estimation (COLLATE)",* IEEE Transactions on Medical Imaging. September 2011.

2. Highly Selective Conference Publications

1. **Andrew J. Asman**, Andrew J. Plassard, Bennett A. Landman. *"Geodesic Learner Fusion or: How We Learned to Stop Worrying and Love Big Data",* Submitted to International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Boston, USA, September 2014
2. Zhoubing Xu, **Andrew J. Asman**, Peter L. Shanahan, Richard G. Abramson, Bennett A. Landman. *"SIMPLE Is a Good Idea (and Better with Context Learning)",* Submitted to International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Boston, USA, September 2014
3. **Andrew J. Asman**, Seth A. Smith, Daniel S. Reich and Bennett A. Landman. *"Robust GM/WM Segmentation of the Spinal Cord with Iterative Non-Local Statistical Fusion"*, In International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Nagoya, Japan, September 2013

4. **Andrew J. Asman** and Bennett A. Landman. "*Non-Local STAPLE: An Intensity-Driven Multi-Atlas Rater Model"*, In International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Nice, France, September 2012
5. Carolyn B. Lauzon, **Andrew J. Asman**, Brian Caffo, Bennett A. Landman. *"Assessment of Bias for MRI Diffusion Tensor Imaging Using SIMEX"*, In International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Toronto, Canada, September 2011
6. **Andrew J. Asman** and Bennett A. Landman. *"Characterizing Spatially Varying Performance to Improve Multi-Atlas Multi-Label Segmentation",* In Proceedings of the 2011 International Conference on Information Processing in Medical Imaging, Irsee, Bavaria, July 2011. Oral Presentation

3. Refereed Conference Publications

1. **Andrew J. Asman**, Alexander S. Dagley, and Bennett A. Landman. *"Statistical label fusion with hierarchical performance models"* In Proceedings of the SPIE Medical Imaging Conference. San Diego, California, February 2014. Oral Presentation. *Best Student Paper Finalist.*
2. Swetasudha Panda, **Andrew J. Asman**, Michael P. DeLisi, Louise A. Mawn, Robert L. Galloway, Bennett A. Landman. *"Robust Optic Nerve Segmentation on Clinically Acquired CT."* In Proceedings of the SPIE Medical Imaging Conference. San Diego, California, February 2014
3. Ryan D. Datteri, **Andrew J. Asman**, Bennett A. Landman, Benoit M. Dawant, Applying the Algorithm *"Assessing Quality Using Image Registration Circuits (AQUIRC) to Multi-Atlas Segmentation".* In Proceedings of the. SPIE, Medical Imaging Conference. San Diego, California, February 2014

4. Zhoubing Xu, Bo Li, Swetasudha Panda, **Andrew J. Asman**, Kristen. L. Merkle, Peter L. Shanahan, Richard G. Abramson, Bennett A. Landman. *"Shape-Constrained Multi-Atlas Segmentation of Spleen in CT."* In Proceedings of the SPIE Medical Imaging Conference. San Diego, California, February 2014
5. **Andrew J. Asman**, Michael P. DeLisi, Louise A. Mawn, Robert L. Galloway, and Bennett A. Landman. *"Robust Non-Local Multi-Atlas Segmentation of the Optic Nerve"* In Proceedings of the SPIE Medical Imaging Conference. Orlando, Florida, February 2013. Oral Presentation.
6. **Andrew J. Asman**, Carolyn B. Lauzon, Bennett A. Landman. *"Robust Inter-Modality Multi-Atlas Segmentation for PACS-based DTI Quality Control"*. In Proceedings of the SPIE Medical Imaging Conference. Orlando, Florida, February 2013. Oral Presentation.
7. Wade M. Allen, Zhoubing Xu, **Andrew J. Asman**, Benjamin K. Poulose, Bennett A. Landman. *"Quantitative Anatomical Labeling of the Anterior Abdominal Wall."* In Proceedings of the SPIE Medical Imaging Conference. Orlando, Florida, February 2013. Oral Presentation.
8. **Andrew J. Asman** and Bennett A. Landman, *"Out-of-Atlas Labeling: A Multi-Atlas Approach to Cancer Segmentation"*, In Proceedings of the 2012 International Symposium on Biomedical Imaging (ISBI). Barcelona, Spain
9. Zhoubing Xu, **Andrew J. Asman**, Eesha Singh, Lola Chambless, Reid Thompson, and Bennett A. Landman, *"Collaborative Labeling of Malignant Glioma",* In Proceedings of the 2012 International Symposium on Biomedical Imaging (ISBI). Barcelona, Spain
10. **Andrew J. Asman** and Bennett A. Landman. *"Simultaneous Segmentation and Statistical Label Fusion."* In Proceedings of the SPIE Medical Imaging Conference. San Diego, California, February 2012. Oral Presentation
11. Zhoubing Xu, **Andrew J. Asman** and Bennett A. Landman. *"Generalized Statistical Label Fusion using Multiple Consensus Levels."* In Proceedings of the SPIE Medical Imaging Conference. San Diego, California, February 2012. Oral Presentation

12. Eesha Singh, Zhoubing Xu, **Andrew J. Asman**, Lola Chambless, Reid Thompson and Bennett A. Landman. *"Collaborative Labeling of Malignant Glioma with WebMILL: A First Look."* In Proceedings of the SPIE Medical Imaging Conference. San Diego, California, February 2012
13. Fangxu Xing, **Andrew J. Asman**, Jerry L. Prince, Bennett A. Landman. *"Finding Seeds for Segmentation Using Statistical Fusion."* In Proceedings of the SPIE Medical Imaging Conference. San Diego, California, February 2012
14. Ryan D. Datteri, **Andrew J. Asman**, Bennett A. Landman, and Benoit M. Dawant. *"Estimation of Registration Accuracy Applied to Multi-Atlas Segmentation"*, In MICCAI: Workshop on Multi-Atlas Labeling and Statistical Fusion, Toronto, Canada, September 2011. Oral Presentation
15. **Andrew J. Asman**, Antong Chen, and Bennett A. Landman. *"On the Application of Human Rater Models to Statistical Fusion in Multi-Atlas Labeling"*, In MICCAI: Workshop on Multi-Atlas Labeling and Statistical Fusion, Toronto, Canada, September 2011. Oral Presentation
16. **Andrew J. Asman**, Andrew G. Scoggins, Jerry L. Prince, Bennett A. Landman. *"Foibles, Follies, and Fusion: Assessment of Statistical Label Fusion Techniques for Web-Based Collaborations using Minimal Training",* In Proceedings of the SPIE Medical Imaging Conference. Lake Buena Vista, Florida, February 2011
17. Joshua A. Stein, **Andrew J. Asman**, Bennett A. Landman. *"Characterizing and Optimizing Rater Performance for Internet-based Collaborative Labeling"*, In Proceedings of the SPIE Medical Imaging Conference. Lake Buena Vista, Florida, February 2011. Oral Presentation
18. **Andrew J. Asman**, Edward E. Rippetoe, and Brian E. Cooper, *"Scanner Characterization for Color Measurement of EP Printed Output",* NIP25: International Conference on Digital Printing Technologies and Digital Fabrication. Louisville, Kentucky, September 2009

4. Conference Publication Abstracts

1. Seth A. Smith, Kanagalingam Sivashakthi, Swetasudha Panda, **Andrew J. Asman**, Sarita Dave, Bennett A. Landman, Blake E. Dewey, Ha Kyu Jeong, Louis A. Mawn. *"Advanced MRI of Optic Nerve Drusen: Preliminary Findings."* North American Neuro-Ophthalmology Society (NANOS), Snowbird, Utah, February 2013
2. Swetasudha Panda, **Andrew J. Asman**, Louise A. Mawn, Bennett A. Landman, Seth A. Smith. *"Robust Segmentation of Clinical Optic Nerve MRI."* International Society for Magnetic Resonance in Medicine, Salt Lake City, UT. April 25, 2013

5. Books / Book Chapters

1. Bennett A Landman, Antong Chen, D. Louis Collins, Pierrick Coupe, Meritxell Bach Cuadra, Ryan D. Datteri, Benoit Dawant, Michal Depa, Simon Duchesne, Simon F. Eskildsen, Vladimir Fonov, Polina Golland, Subrahmanyam Gorthi, Abdelkarim S. Allal, **Andrew J. Asman**, Josephine Barnes, Floris F. Berendsen, M. Jorge Cardoso, Arion F. Chatziioannou, Simon Warfield, *MICCAI 2011 Workshop on Multi-Atlas Labeling and Statistical Fusion*, CreateSpace Independent Publishing Platform (August 26, 2012)
2. Bennett A Landman, Annemie Ribbens, Blake Lucas, Christos Davatzikos, Brian Avants, Christian Ledig, Da Ma, Daniel Rueckert, Dirk Vandermeulen, Frederik Maes, Guray Erus, Jiahui Wang, Holly Holmes, Hongzhi Wang, Jimit Doshi, Joe Kornegay , Jose Manjon, Alexander Hammers, Alireza Akhondi-Asl, **Andrew J. Asman**, Simon K Warfield, *MICCAI 2012 Workshop on Multi-Atlas Labeling*, CreateSpace Independent Publishing Platform (September 22, 2011)

APPENDIX B

BIOGRAPHY

Andrew J. Asman was born in Ft. Wright, Kentucky in 1987. He received dual B.S. degrees in electrical engineering and computer engineering from the University of Kentucky, Lexington, Kentucky, in 2010, and a Ph.D. degree in electrical engineering from Vanderbilt University, Nashville, Tennessee, in 2014.

As an undergraduate, he worked at Lexmark International, Inc., as a student software developer in the Colorscience and Imaging Department. Simultaneously, he worked as a research assistant at the University of Kentucky where his work focused on audio rendering, sound source localization and cocktail party problems. As a graduate student, his work focused on supervised learning and statistical modeling of rater performance for multi-atlas segmentation of medical images. Additionally, his research interests include detecting imaging abnormalities and anomalies and leveraging structural shape/appearance variability to form groupwise consistent models of anatomy.

Dr. Asman is a member of the National Society of Collegiate Scholars (NSCS), the Institute of Electrical and Electronics Engineers (IEEE), .International Society for Optics and Photonics (SPIE), and the Medical Image Computing and Computer Assisted Intervention (MICCAI).

REFERENCES

1. Crespo-Facorro, B., et al., *Human frontal cortex: an MRI-based parcellation method.* NeuroImage, 1999. **10**(5): p. 500-519.
2. Falcão, A.X., et al., *User-steered image segmentation paradigms: Live wire and live lane.* Graphical models and image processing, 1998. **60**(4): p. 233-260.
3. Tsang, O., et al., *Comparison of tissue segmentation algorithms in neuroimage analysis software tools.* Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference, 2008. **2008**: p. 3924-8.
4. Yushkevich, P.A., et al., *User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability.* NeuroImage, 2006. **31**(3): p. 1116-1128.
5. Ashton, E.A., et al., *Accuracy and reproducibility of manual and semiautomated quantification of MS lesions by MRI.* Journal of Magnetic Resonance Imaging, 2003. **17**(3): p. 300-308.
6. Joe, B.N., et al., *Brain Tumor Volume Measurement: Comparison of Manual and Semiautomated Methods1*. Radiology, 1999. **212**(3): p. 811-816.
7. Kaus, M., et al. *Segmentation of meningiomas and low grade gliomas in MRI*. in *Medical Image Computing and Computer-Assisted Intervention–MICCAI'99*. 1999. Springer.
8. Warfield, S.K., K.H. Zou, and W.M. Wells, *Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation.* IEEE Transactions on Medical Imaging, 2004. **23**(7): p. 903-921.
9. Rohlfing, T., D.B. Russakoff, and C.R. Maurer, *Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation.* IEEE Transactions on Medical Imaging, 2004. **23**(8): p. 983-994.

10. Landman, B.A., et al., *Foibles, follies, and fusion: Web-based collaboration for medical image labeling.* NeuroImage, 2011.
11. Landman, B.A., et al., *Robust Statistical Fusion of Image Labels.* IEEE Transactions on Medical Imaging, 2011. **31**(2): p. 512-522.
12. Asman, A.J., et al. *Foibles, follies, and fusion: assessment of statistical label fusion techniques for web-based collaborations using minimal training*. in *SPIE Medical Imaging*. 2011. International Society for Optics and Photonics.
13. Singh, E., et al. *Collaborative labeling of malignant glioma with WebMILL: a first look*. in *SPIE Medical Imaging*. 2012.
14. Xu, Z., et al. *Collaborative labeling of malignant glioma*. in *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*. 2012. IEEE.
15. Xu, Z., et al., *Segmentation of malignant gliomas through remote collaboration and statistical fusion.* Medical Physics, 2012. **39**(10): p. 5981.
16. Bryan, F.W., et al., *Self-assessed performance improves statistical fusion of image labels.* Medical Physics, 2014. **41**(3): p. 031903.
17. Wells III, W., et al., *Adaptive segmentation of MRI data.* Medical Imaging, IEEE Transactions on, 1996. **15**(4): p. 429-442.
18. Van Leemput, K., et al., *Automated model-based tissue classification of MR images of the brain.* Medical Imaging, IEEE Transactions on, 1999. **18**(10): p. 897-908.
19. Ashburner, J. and K.J. Friston, *Unified segmentation.* NeuroImage, 2005. **26**(3): p. 839-851.
20. Kapur, T., et al., *Segmentation of brain tissue from magnetic resonance images.* Medical Image Analysis, 1996. **1**(2): p. 109-127.
21. Tanabe, J.L., et al., *Tissue segmentation of the brain in Alzheimer disease.* American Journal of Neuroradiology, 1997. **18**(1): p. 115-123.
22. Cohen, G., et al., *Segmentation techniques for the classification of brain tissue using magnetic resonance imaging.* Psychiatry Research: Neuroimaging, 1992. **45**(1): p. 33-51.

23. Pham, D.L. and J.L. Prince, *Adaptive fuzzy segmentation of magnetic resonance images.* Medical Imaging, IEEE Transactions on, 1999. **18**(9): p. 737-752.

24. Cardoso, M.J., et al., *LoAd: A locally adaptive cortical segmentation algorithm.* NeuroImage, 2011. **56**(3): p. 1386-1397.

25. Fischl, B., et al., *Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain.* Neuron, 2002. **33**(3): p. 341-355.

26. Heckemann, R.A., et al., *Automatic anatomical brain MRI segmentation combining label propagation and decision fusion.* NeuroImage, 2006. **33**(1): p. 115-126.

27. Yeo, B.T.T., et al., *Effects of registration regularization and atlas sharpness on segmentation accuracy*. Medical Image Analysis, 2008. **12**(5): p. 603-615.

28. Gee, J.C., M. Reivich, and R. Bajcsy, *Elastically deforming a three-dimensional atlas to match anatomical brain images.* Journal of Computer Assisted Tomography, 1993. **17**(2): p. 225-236.

29. Noble, J.H. and B.M. Dawant, *An atlas-navigated optimal medial axis and deformable model algorithm (NOMAD) for the segmentation of the optic nerves and chiasm in MR and CT images.* Medical Image Analysis, 2011. **15**(6): p. 877-884.

30. Lee, J.M., et al., *Evaluation of automated and semi-automated skull-stripping algorithms using similarity index and segmentation error.* Computers in Biology and Medicine, 2003. **33**(6): p. 495-507.

31. Huang, T.C., et al., *Semi-automated CT segmentation using optic flow and Fourier interpolation techniques.* Computer Methods and Programs in Biomedicine, 2006. **84**(2-3): p. 124-134.

32. Noble, J. and B. Dawant, *A new approach for tubular structure modeling and segmentation using graph-based techniques.* Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011, 2011: p. 305-312.

33. Collins, D.L., et al., *Automatic 3-D model-based neuroanatomical segmentation.* Human Brain Mapping, 1995. **3**(3): p. 190-208.

34. Dawant, B.M., et al., *Automatic 3-D segmentation of internal structures of the head in MR images using a combination of similarity and free-form transformations. I. Methodology and validation on normal subjects.* Medical Imaging, IEEE Transactions on, 1999. **18**(10): p. 909-916.

35. Collins, D.L. and A.C. Evans, *Animal: validation and applications of nonlinear registration-based segmentation.* International Journal of Pattern Recognition and Artificial Intelligence, 1997. **11**(08): p. 1271-1294.

36. Avants, B., et al., *Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain.* Medical Image Analysis, 2008. **12**(1): p. 26-41.

37. Ardekani, B.A., et al., *A fully automatic multimodality image registration algorithm.* Journal of Computer Assisted Tomography, 1995. **19**(4): p. 615.

38. Rohde, G.K., A. Aldroubi, and B.M. Dawant, *The adaptive bases algorithm for intensity-based nonrigid image registration.* Medical Imaging, IEEE Transactions on, 2003. **22**(11): p. 1470-1479.

39. Klein, A., et al., *Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration.* NeuroImage, 2009. **46**(3): p. 786-802.

40. Guimond, A., J. Meunier, and J.P. Thirion, *Average brain models: A convergence study.* Computer vision and image understanding, 2000. **77**(2): p. 192-210.

41. Joshi, S., et al., *Unbiased diffeomorphic atlas construction for computational anatomy.* NeuroImage, 2004. **23**: p. S151-S160.

42. Fonov, V., et al., *Unbiased average age-appropriate atlases for pediatric studies.* NeuroImage, 2011. **54**(1): p. 313.

43. Ericsson, A., P. Aljabar, and D. Rueckert. *Construction of a patient-specific atlas of the brain: Application to normal aging*. 2008. IEEE.

44. Commowick, O., S. Warfield, and G. Malandain, *Using Frankenstein's creature paradigm to build a patient specific atlas.* Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009, 2009: p. 993-1000.

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

46. Asman, A.J. and B.A. Landman. *Simultaneous Segmentation and Statistical Label Fusion*. in *SPIE Medical Imaging*. 2012. San Diego, CA.

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

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

49. Asman, A. and B. Landman. *Characterizing spatially varying performance to improve multi-atlas multi-label segmentation*. in *Information Processing in Medical Imaging (IPMI)*. 2011. Springer.

50. Asman, A. and B. Landman, *Robust Statistical Label Fusion through Consensus Level, Labeler Accuracy and Truth Estimation (COLLATE).* Medical Imaging, IEEE Transactions on, 2011. **30**(10): p. 1779-1794.

51. Asman, A.J. and B.A. Landman. *Non-Local STAPLE: An Intensity-Driven Multi-Atlas Rater Model*. in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2012. Nice, France: Springer.

52. Asman, A.J. and B.A. Landman, *Formulating Spatially Varying Performance in the Statistical Fusion Framework.* IEEE Transactions on Medical Imaging, 2012. **31**(6): p. 1326 - 1336.

53. Cardoso, M.J., et al. *Locally Ranked STAPLE for template based segmentation propagation*. in *MICCAI Workshop on Multi-Atlas Labeling and Statistical Fusion*. 2011.

54. Commowick, O., A. Akhondi-Asl, and S.K. Warfield, *Estimating A Reference Standard Segmentation with Spatially Varying Performance Parameters: Local MAP STAPLE.* IEEE transactions on medical imaging, 2012. **31**(8): p. 1593-1606.

55. Commowick, O. and S. Warfield, *Incorporating priors on expert performance parameters for segmentation validation and label fusion: a maximum a posteriori STAPLE.* Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010, 2010: p. 25-32.

56. Coupé, P., et al., *Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation.* NeuroImage, 2011. **54**(2): p. 940-954.

57. Isgum, I., et al., *Multi-atlas-based segmentation with local decision fusion—Application to cardiac and aortic segmentation in CT scans.* Medical Imaging, IEEE Transactions on, 2009. **28**(7): p. 1000-1010.

58. Langerak, T.R., et al., *Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE).* Medical Imaging, IEEE Transactions on, 2010. **29**(12): p. 2000-2008.

59. Sabuncu, M.R., et al., *A generative model for image segmentation based on label fusion.* IEEE Transactions on Medical Imaging, 2010. **29**(10): p. 1714-1729.

60. Wang, H., et al. *Optimal weights for multi-atlas label fusion*. in *Information Processing in Medical Imaging (IPMI)*. 2011. Springer.

61. Wang, H., et al., *Multi-Atlas Segmentation with Joint Label Fusion.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012. **35**(3): p. 611-623.

62. Rohlfing, T. and C.R. Maurer, *Shape-based averaging.* Image Processing, IEEE Transactions on, 2007. **16**(1): p. 153-161.

63. Aljabar, P., et al., *Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy*. NeuroImage, 2009. **46**(3): p. 726-738.

64. Rohlfing, T., et al., *Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains.* NeuroImage, 2004. **21**(4): p. 1428-1442.

65. Wolz, R., et al., *LEAP: Learning embeddings for atlas propagation.* NeuroImage, 2010. **49**(2): p. 1316-1325.

66. Cardoso, M., et al., *Geodesic Information Flows.* Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012, 2012: p. 262-270.

67. Weisenfeld, N. and S. Warfield. *Learning likelihoods for labeling (L3): a general multi-classifier segmentation algorithm*. in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2011. Springer.

68. Lotjonen, J.M.P., et al., *Fast and robust multi-atlas segmentation of brain magnetic resonance images.* NeuroImage, 2010. **49**(3): p. 2352-2365.

69. Ledig, C., et al. *Multi-class brain segmentation using atlas propagation and EM-based refinement*. in *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*. 2012. IEEE.

70. Wang, H., et al., *A learning-based wrapper method to correct systematic errors in automatic image segmentation: consistently improved performance in hippocampus, cortex and brain segmentation.* NeuroImage, 2011. **55**(3): p. 968-985.

71. Klein, A. and J. Hirsch, *Mindboggle: a scatterbrained approach to automate brain labeling.* NeuroImage, 2005. **24**(2): p. 261.

72. Chen, A., et al., *Evaluation of multi atlas-based approaches for the segmentation of the thyroid gland in IMRT head-and-neck CT images.* Physics in Medicine and Biology, 2011. **57**: p. 93-11.

73. van Rikxoort, E.M., et al., *Adaptive local multi-atlas segmentation: application to the heart and the caudate nucleus.* Medical Image Analysis, 2010. **14**(1): p. 39-49.

74. Bai, W., et al., *A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: application to cardiac MR images.* IEEE transactions on medical imaging, 2013. **32**(7): p. 1302-15.

75. Wolz, R., et al., *Multi-organ Abdominal CT Segmentation Using Hierarchically Weighted Subject-Specific Atlases.* Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012, 2012: p. 10-17.

76. Criminisi, A., et al., *Regression forests for efficient anatomy detection and localization in computed tomography scans.* Medical Image Analysis, 2013. **17**(8): p. 1293-303.

77. Xu, Z., A.J. Asman, and B.A. Landman. *Generalized Statistical Label Fusion using Multiple Consensus Levels*. in *SPIE Medical Imaging*. 2012. San Diego, CA.

78. Asman, A.J. and B.A. Landman, *Non-Local Statistical Label Fusion for Multi-Atlas Segmentation.* Medical Image Analysis, 2012. **17**(2): p. 194-208.

79. Asman, A.J., A.S. Dagley, and B.A. Landman. *Statistical label fusion with hierarchical performance models*. in *SPIE Medical Imaging*. 2014. San Diego, CA: International Society for Optics and Photonics.

80. Talairach, J. and P. Tournoux, *Co-planar stereotaxic atlas of the human brain. 3-Dimensional proportional system: an approach to cerebral imaging.* 1988.

81. Evans, A., et al., *MRI-PET correlation in three dimensions using a volume-of-interest (VOI) atlas.* Journal of Cerebral Blood Flow and Metabolism, 1991. **11**: p. A69-A78.

82. Evans, A., D. Collins, and B. Milner. *An MRI-based stereotactic atlas from 250 young normal subjects*. in *Soc. neurosci. abstr*. 1992.

83. Evans, A.C., et al., *Anatomical mapping of functional activation in stereotactic coordinate space.* NeuroImage, 1992. **1**(1): p. 43-53.

84. Greitz, T., et al., *A computerized brain atlas: construction, anatomical content, and some applications.* Journal of Computer Assisted Tomography, 1991. **15**(1): p. 26.

85. Seitz, R., et al., *Accuracy and precision of the computerized brain atlas programme for localization and quantification in positron emission tomography.* Journal of Cerebral Blood Flow and Metabolism, 1990. **10**(4): p. 443-457.

86. Evans, A.C., et al. *3D statistical neuroanatomical models from 305 MRI volumes*. in *Nuclear Science Symposium and Medical Imaging Conference, 1993., 1993 IEEE Conference Record*. 1993. IEEE.

87. Kamber, M., et al., *Model-based 3-D segmentation of multiple sclerosis lesions in magnetic resonance brain images.* Medical Imaging, IEEE Transactions on, 1995. **14**(3): p. 442-453.

88. Mazziotta, J.C., et al., *A probabilistic atlas of the human brain: theory and rationale for its development. The International Consortium for Brain Mapping (ICBM).* NeuroImage, 1995. **2**(2): p. 89.

89. Holmes, C.J., et al., *Enhancement of MR images using registration for signal averaging.* Journal of Computer Assisted Tomography, 1998. **22**(2): p. 324-333.

90. Woolrich, M.W., et al., *Bayesian analysis of neuroimaging data in FSL.* NeuroImage, 2009. **45**(1 Suppl): p. S173-S186.

91. Friston, K.J. and J. Ashburner, *Statistical parametric mapping.* Functional neuroimaging: Technical foundations, 1994: p. 79-93.

92. Maintz, J.B. and M.A. Viergever, *A survey of medical image registration.* Medical Image Analysis, 1998. **2**(1): p. 1-36.

93. Maes, F., et al., *Multimodality image registration by maximization of mutual information.* IEEE transactions on medical imaging, 1997. **16**(2): p. 187-98.

94. Ourselin, S., et al. *Block matching: A general framework to improve robustness of rigid registration of medical images*. in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2000*. 2000. Springer.

95. Fitzpatrick, J.M., J.B. West, and C.R. Maurer Jr, *Predicting error in rigid-body point-based registration.* Medical Imaging, IEEE Transactions on, 1998. **17**(5): p. 694-702.

96. West, J.B., et al., *Fiducial point placement and the accuracy of point-based, rigid body registration.* Neurosurgery, 2001. **48**(4): p. 810-817.

97. Fitzpatrick, J.M. and J.B. West, *The distribution of target registration error in rigid-body point-based registration.* Medical Imaging, IEEE Transactions on, 2001. **20**(9): p. 917-927.

98. Jenkinson, M. and S. Smith, *A global optimisation method for robust affine registration of brain images.* Medical Image Analysis, 2001. **5**(2): p. 143-156.

99. Denton, E.R.E., et al., *Comparison and evaluation of rigid, affine, and nonrigid registration of breast MR images.* Journal of Computer Assisted Tomography, 1999. **23**(5): p. 800-805.

100. Feldmar, J. and N. Ayache, *Rigid, affine and locally affine registration of free-form surfaces.* International journal of computer vision, 1996. **18**(2): p. 99-119.

101. Ourselin, S., et al., *Reconstructing a 3D structure from serial histological sections.* Image and vision computing, 2001. **19**(1): p. 25-31.

102. Rueckert, D., et al., *Nonrigid registration using free-form deformations: application to breast MR images.* Medical Imaging, IEEE Transactions on, 1999. **18**(8): p. 712-721.

103. Fiebich, M., et al., *Automatic bone segmentation technique for CT angiographic studies.* Journal of Computer Assisted Tomography, 1999. **23**(1): p. 155-161.

104. Asman, A.J., et al. *Robust non-local multi-atlas segmentation of the optic nerve*. in *SPIE Medical Imaging*. 2013. International Society for Optics and Photonics.

105. Cootes, T.F., et al., *Active shape models-their training and application.* Computer vision and image understanding, 1995. **61**(1): p. 38-59.

106. Cootes, T.F., G.J. Edwards, and C.J. Taylor, *Active appearance models.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2001. **23**(6): p. 681-685.

107. Balci, S.K., et al. *Free-form B-spline deformation model for groupwise registration*. in *Medical image computing and computer-assisted intervention: MICCAI... International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2007. NIH Public Access.

108. Kovacevic, N., et al., *Deformation based representation of groupwise average and variability.* Medical Image Computing and Computer-Assisted Intervention–MICCAI 2004, 2004: p. 615-622.

109. Depa, M., et al. *Towards efficient label fusion by pre-alignment of training data*. in *Proc. MICCAI Workshop on Multi-atlas Labeling and Statistical Fusion*. 2011.

110. Heckemann, R.A., et al., *Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation.* NeuroImage, 2010. **51**(1): p. 221.

111. Shen, D. and C. Davatzikos, *HAMMER: hierarchical attribute matching mechanism for elastic registration.* Medical Imaging, IEEE Transactions on, 2002. **21**(11): p. 1421-1439.

112. Kearns, M.J. and L.G. Valiant, *Learning Boolean formulae or finite automata is as hard as factoring*1988: Harvard University, Center for Research in Computing Technology, Aiken Computation Laboratory.

113. Schapire, R.E., *The strength of weak learnability.* Machine learning, 1990. **5**(2): p. 197-227.

114. Freund, Y. and R. Schapire. *A desicion-theoretic generalization of on-line learning and an application to boosting*. in *Computational learning theory*. 1995. Springer.

115. Parzen, E., *On estimation of a probability density function and mode.* The annals of mathematical statistics, 1962. **33**(3): p. 1065-1076.

116. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society. Series B (Methodological), 1977: p. 1-38.

117. Bellman, R., *Dynamic programming and Lagrange multipliers.* Proceedings of the National Academy of Sciences of the United States of America, 1956. **42**(10): p. 767.

118. Cardoso, M.J., et al., *STEPS: Similarity and Truth Estimation for Propagated Segmentations and its application to hippocampal segmentation and brain parcelation.* Medical Image Analysis, 2013. **17**(6): p. 671-684.

119. Collewet, G., M. Strzelecki, and F. Mariette, *Influence of MRI acquisition protocols and image intensity normalization methods on texture classification.* Magnetic Resonance Imaging, 2004. **22**(1): p. 81-91.

120. Madabhushi, A. and J.K. Udupa, *Interplay between intensity standardization and inhomogeneity correction in MR image processing.* Medical Imaging, IEEE Transactions on, 2005. **24**(5): p. 561-576.

121. Nyu, L.G. and J.K. Udupa, *On standardizing the MR image intensity scale.* Image, 1999. **1081**.

122. Buades, A., B. Coll, and J.M. Morel. *A non-local algorithm for image denoising*. in *Computer Vision and Pattern Recognition (CVPR)*. 2005. IEEE.

123. Coupé, P., P. Yger, and C. Barillot, *Fast non local means denoising for 3D MR images.* Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006, 2006: p. 33-40.

124. Kervrann, C., J. Boulanger, and P. Coupé. *Bayesian non-local means filter, image redundancy and adaptive dictionaries for noise removal*. 2007. Springer-Verlag.

125. Manjón, J.V., et al., *MRI denoising using non-local means.* Medical Image Analysis, 2008. **12**(4): p. 514-523.

126. Liu, Y.L., et al., *A robust and fast non-local means algorithm for image denoising.* Journal of Computer Science and Technology, 2008. **23**(2): p. 270-279.

127. Van De Ville, D. and M. Kocher, *SURE-based non-local means.* Signal Processing Letters, IEEE, 2009. **16**(11): p. 973-976.

128. Besag, J., *On the statistical analysis of dirty pictures.* Journal of the Royal Statistical Society. Series B (Methodological), 1986: p. 259-302.

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

130. Held, K., et al., *Markov random field segmentation of brain MR images.* Medical Imaging, IEEE Transactions on, 1997. **16**(6): p. 878-886.

131. Asman, A.J. and B.A. Landman. *Out-of-atlas labeling: A multi-atlas approach to cancer segmentation*. in *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*. 2012. IEEE.

132. Asman, A.J., et al. *Robust GM/WM Segmentation of the Spinal Cord with Iterative Non-Local Statistical Fusion*. in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2013. Nagoya, Japan: Springer.

133. Asman, A.J., et al., *Groupwise multi-atlas segmentation of the spinal cord's internal structure.* Medical Image Analysis, 2014. **18**(3): p. 460-471.

134. Udupa, J.K., et al., *A framework for evaluating image segmentation algorithms.* Computerized Medical Imaging and Graphics, 2006. **30**(2): p. 75-87.

135. Warfield, S.K., K.H. Zou, and W.M. Wells, *Validation of image segmentation by estimating rater bias and variance.* Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2008. **366**(1874): p. 2361-2375.

136. Xing, F., et al. *Statistical fusion of continuous labels: identification of cardiac landmarks*. in *SPIE Medical Imaging*. 2011. International Society for Optics and Photonics.

137. Collins, D.L., *3D Model-based segmentation of individual brain structures from magnetic resonance imaging data*, 1994, McGill University.

138. McLachlan, G.J. and T. Krishnan, *The EM algorithm and extensions*. Vol. 382. 2007: Wiley-Interscience.

139. Moon, T.K., *The expectation-maximization algorithm.* Signal Processing Magazine, IEEE, 1996. **13**(6): p. 47-60.

140. Dice, L.R., *Measures of the amount of ecologic association between species.* Ecology, 1945. **26**(3): p. 297-302.

141. Jaccard, P., *The distribution of the flora in the alpine zone.* New Phytologist, 2006. **11**(2): p. 37-50.

142. Landman, B.A., J.A. Bogovic, and J.L. Prince. *Simultaneous truth and performance level estimation with incomplete, over-complete, and ancillary data*. in *Proceedings-Society of Photo-Optical Instrumentation Engineers*. 2010. NIH Public Access.

143. Bogovic, J., et al. *Statistical fusion of surface labels provided by multiple raters, over-complete, and ancillary data*. in *SPIE Medical Imaging Conference*. 2010.

144. Rohlfing, T., D. Russakoff, and C. Maurer, *Performance-Based Classifier Combination in Atlas-Based Image Segmentation Using Expectation-Maximization Parameter Estimation.* IEEE Transactions on Medical Imaging, 2004. **23**: p. 983-994.

145. Marcus, D.S., et al., *Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults.* Journal of Cognitive Neuroscience, 2007. **19**(9): p. 1498-1507.

146. Roy, S., A. Carass, and J.L. Prince. *Synthesizing MR contrast and resolution through a patch matching technique*. 2010. NIH Public Access.

147. Sun, J. and M.F. Tappen. *Learning non-local range Markov Random field for image restoration*. 2011. IEEE.

148. Roy, S., et al. *MR contrast synthesis for lesion segmentation*. 2010. IEEE.

149. Huttenlocher, D.P., G.A. Klanderman, and W.J. Rucklidge, *Comparing images using the Hausdorff distance.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1993. **15**(9): p. 850-863.

150. Najman, L. and M. Schmitt, *Geodesic saliency of watershed contours and hierarchical segmentation.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996. **18**(12): p. 1163-1173.

151. Beucher, S., *Watershed, hierarchical segmentation and waterfall algorithm*, in *Mathematical morphology and its applications to image processing*1994, Springer. p. 69-76.

152. Lucas, B.C., et al., *The Java Image Science Toolkit (JIST) for rapid prototyping and publishing of neuroimaging software.* Neuroinformatics, 2010. **8**(1): p. 5-17.

153. Wilcoxon, F., *Individual comparisons by ranking methods.* Biometrics bulletin, 1945. **1**(6): p. 80-83.

154. Klein, A., et al. *Open labels: online feedback for a public resource of manually labeled brain images*. in *16th Annual Meeting for the Organization of Human Brain Mapping*. 2010.

155. Avants, B.B., et al., *A reproducible evaluation of ANTs similarity metric performance in brain image registration.* NeuroImage, 2011. **54**(3): p. 2033-44.

156. Prastawa, M., et al., *Automatic brain tumor segmentation by subject specific modification of atlas priors.* Academic Radiology, 2003. **10**(12): p. 1341-1348.

157. Gering, D., W. Grimson, and R. Kikinis, *Recognizing deviations from normalcy for brain tumor segmentation.* Medical Image Computing and Computer-Assisted Intervention—MICCAI 2002, 2002: p. 388-395.

158. Corso, J.J., et al., *Efficient multilevel brain tumor segmentation with integrated bayesian model classification*. Medical Imaging, IEEE Transactions on, 2008. **27**(5): p. 629-640.

159. van Ginneken, B., et al., *Automatic detection of abnormalities in chest radiographs using local texture analysis*. Medical Imaging, IEEE Transactions on, 2002. **21**(2): p. 139-149.

160. Krishnan, S., et al. *Intestinal abnormality detection from endoscopic images*. in *Engineering in Medicine and Biology Society (EMBS)*. 1998. Hong Kong, China: IEEE.

161. Prastawa, M., et al., *A brain tumor segmentation framework based on outlier detection.* Medical Image Analysis, 2004. **8**(3): p. 275-283.

162. Gholipour, A., et al., *Multi-atlas multi-shape segmentation of fetal brain MRI for volumetric and morphometric analysis of ventriculomegaly.* NeuroImage, 2012.

163. Gensheimer, M., et al. *Automatic delineation of the optic nerves and chiasm on CT images*. in *SPIE Medical Imaging*. 2007.

164. Amari, S., A. Cichocki, and H.H. Yang, *A new learning algorithm for blind signal separation.* Advances in neural information processing systems, 1996: p. 757-763.

165. McAuliffe, M.J., et al. *Medical image processing, analysis and visualization in clinical research*. in *Computer-Based Medical Systems, 2001. CBMS 2001. Proceedings. 14th IEEE Symposium on*. 2001. IEEE.

166. Dietz, V. and A. Curt, *Neurological aspects of spinal-cord repair: promises and challenges.* The Lancet Neurology, 2006. **5**(8): p. 688-694.

167. Bede, P., et al., *Spinal cord markers in ALS: diagnostic and biomarker considerations.* Amyotrophic lateral sclerosis : official publication of the World Federation of Neurology Research Group on Motor Neuron Diseases, 2012. **13**(5): p. 407-15.

168. Bede, P., et al., *Grey matter correlates of clinical variables in amyotrophic lateral sclerosis (ALS): a neuroimaging study of ALS motor phenotype heterogeneity and cortical focality.* Journal of neurology, neurosurgery, and psychiatry, 2013. **84**(7): p. 766-73.

169. Wingerchuk, D.M., et al., *The spectrum of neuromyelitis optica.* Lancet neurology, 2007. **6**(9): p. 805-15.

170. Yiannakas, M., et al., *Feasibility of Grey Matter and White Matter Segmentation of the Upper Cervical Cord In Vivo: A pilot study with application to Magnetisation Transfer Measurements.* NeuroImage, 2012. **63**(3): p. 1054-1059.

171. Gilmore, C.P., et al., *Spinal cord gray matter demyelination in multiple sclerosis—a novel pattern of residual plaque morphology*. Brain Pathology, 2006. **16**(3): p. 202-208.

172. Jarius, S. and B. Wildemann, *AQP4 antibodies in neuromyelitis optica: diagnostic and pathogenetic relevance.* Nature Reviews Neurology, 2010. **6**(7): p. 383-392.

173. Chen, M., et al. *Topology preserving automatic segmentation of the spinal cord in magnetic resonance images*. in *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*. 2011. IEEE.

174. McIntosh, C. and G. Hamarneh, *Spinal crawlers: deformable organisms for spinal cord segmentation and analysis.* Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention, 2006. **9**(Pt 1): p. 808-15.

175. Carballido-Gamio, J., S.J. Belongie, and S. Majumdar, *Normalized cuts in 3-D for spinal MRI segmentation.* IEEE transactions on medical imaging, 2004. **23**(1): p. 36-44.

176. Huang, S.H., et al., *Learning-based vertebra detection and iterative normalized-cut segmentation for spinal MRI.* IEEE transactions on medical imaging, 2009. **28**(10): p. 1595-605.

177. Ma, J., et al., *Hierarchical segmentation and identification of thoracic vertebra using learning-based edge detection and coarse-to-fine deformable model.* Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention, 2010. **13**(Pt 1): p. 19-27.

178. Horsfield, M.A., et al., *Rapid semi-automatic segmentation of the spinal cord from magnetic resonance images: Application in multiple sclerosis.* NeuroImage, 2010. **50**(2): p. 446-455.

179. Kaminsky, J., et al., *Specially adapted interactive tools for an improved 3D-segmentation of the spine.* Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society, 2004. **28**(3): p. 119-27.

180. Zackowski, K.M., et al., *Sensorimotor dysfunction in multiple sclerosis and column-specific magnetization transfer-imaging abnormalities in the spinal cord.* Brain : a journal of neurology, 2009. **132**(Pt 5): p. 1200-9.

181. Farrell, J.A., et al., *High b-value q-space diffusion-weighted MRI of the human cervical spinal cord in vivo: feasibility and application to multiple sclerosis.* Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine, 2008. **59**(5): p. 1079-89.

182. Smith, S.A., et al., *Reproducibility of tract-specific magnetization transfer and diffusion tensor imaging in the cervical spinal cord at 3 tesla.* NMR in Biomedicine, 2010. **23**(2): p. 207-17.

183. Ozturk, A., et al., *Axial 3D gradient-echo imaging for improved multiple sclerosis lesion detection in the cervical spinal cord at 3T.* Neuroradiology, 2013. **55**(4): p. 431-9.

184. Mikulis, D.J., et al., *Oscillatory motion of the normal cervical spinal cord.* Radiology, 1994. **192**(1): p. 117-21.

185. Karpova, A., et al., *Reliability of quantitative magnetic resonance imaging methods in the assessment of spinal canal stenosis and cord compression in cervical myelopathy.* Spine, 2013. **38**(3): p. 245-52.

186. Hinks, R.S. and R.M. Quencer, *Motion artifacts in brain and spine MR.* Radiologic Clinics of North America, 1988. **26**(4): p. 737-53.

187. Smith, S.A., et al., *Measurement of T1 and T2 in the cervical spinal cord at 3 tesla.* Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine, 2008. **60**(1): p. 213-9.

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

189. Balci, S.K., et al., *Free-Form B-spline Deformation Model for Groupwise Registration.* Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention, 2007. **10**(WS): p. 23-30.

190. Bhatia, K.K., et al., *Similarity metrics for groupwise non-rigid registration.* Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention, 2007. **10**(Pt 2): p. 544-52.

191. Cao, Y., et al., *Segmenting images by combining selected atlases on manifold.* Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention, 2011. **14**(Pt 3): p. 272-9.

192. Turk, M. and A. Pentland, *Eigenfaces for recognition.* Journal of Cognitive Neuroscience, 1991. **3**(1): p. 71-86.

193. Weisenfeld, N.I. and S.K. Warfield, *Automatic segmentation of newborn brain MRI.* NeuroImage, 2009. **47**(2): p. 564-72.

194. Jolliffe, I.T., *Principal component analysis*. Vol. 487. 1986: Springer-Verlag New York.

195. Lagarias, J.C., et al., *Convergence properties of the Nelder--Mead simplex method in low dimensions.* SIAM Journal on Optimization, 1998. **9**(1): p. 112-147.

196. Landman, B.A., et al., *MICCAI 2012 Workshop on Multi-Atlas Labeling*. Vol. 2. 2012: CreateSpace Independent Publishing Platform. 164.

197. Avants, B.B., N. Tustison, and G. Song, *Advanced Normalization Tools (ANTS).* Insight Journal, 2009.

198. Zhang, J., *The mean field theory in EM procedures for Markov random fields.* Signal Processing, IEEE Transactions on, 1992. **40**(10): p. 2570-2583.

199. Pitiot, A., et al., *Piecewise affine registration of biological images for volume reconstruction.* Medical Image Analysis, 2006. **10**(3): p. 465-483.

200. Commowick, O., N. Wiest-Daesslé, and S. Prima, *Automated diffeomorphic registration of anatomical structures with rigid parts: Application to dynamic cervical MRI*. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012, 2012: p. 163-170.

201. Gerber, S., et al., *Manifold modeling for brain population analysis.* Medical Image Analysis, 2010. **14**(5): p. 643-53.

202. Panda, S., et al. *Robust optic nerve segmentation on clinically acquired CT*. in *SPIE Medical Imaging*. 2014. San Diego, CA: International Society for Optics and Photonics.