Multi-Modal Image Registration and Atlas Formation

by
Peter Jonathan Lorenzen

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Computer Science.

Chapel Hill
2006

Approved by
Advisor: Sarang C. Joshi, D.Sc.
Reader: Guido Gerig, Ph.D.
Reader: Stephen M. Pizer, Ph.D.
Reader: Elizabeth Bullitt, M.D.
Reader: Andrew B. Nobel, Ph.D.

Abstract
PETER JONATHAN LORENZEN: Multi-Modal Image Registration and Atlas Formation.
(Under the direction of Sarang C. Joshi, D.Sc..)

Medical images of human anatomy can be produced from a wide range of sensor technologies and imaging techniques resulting in a diverse array of imaging modalities, such as magnetic resonance and computed tomography. The physical properties of the image acquisition process for different modalities elicit different tissue structures. Images from multiple modalities provide complementary information about underlying anatomical structure. Understanding anatomical variability is often important in studying disparate population groups and typically requires robust dense image registration. Traditional image registration methods involve finding a mapping between two scalar images. Such methods do not exploit the complementary information provided by sets of multi-modal images.

This dissertation presents a Bayesian framework for generating inter-subject large deformation transformations between two multi-modal image sets of the brain. The estimated transformations are generated using maximal information about the underlying neuroanatomy present in each of the different modalities. This modality independent registration framework is achieved by jointly estimating the posterior probabilities associated with the multi-modal image sets and the high-dimensional registration transformations relating these posteriors. To maximally use the information present in all the modalities for registration, Kullback-Leibler divergence between the estimated posteriors is minimized. This framework is extended to large deformation multi-class posterior atlas estimation. The method generates a representative anatomical template from an arbitrary number of topologically similar multi-modal image sets. The generated atlas is the class posterior that requires the least amount of deformation energy to be transformed into every class posterior (each characterizing a multi-modal image set). This method is computationally practical in that computation times grows linearly with the number of image sets.

The multi-class posterior atlas formation method is applied to a database of multi-modal images from ninety-five adult brains as part of a healthy aging study to produce 4D spatiotemporal atlases for the female and male subpopulations. The stability of the

atlases is evaluated based on the entropy of their class posteriors. Global volumetric trends and local volumetric change are evaluated. This multi-modal framework has potential applications in many natural multi-modal imaging environments.

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Sarang Joshi for his guidance, motivation, and support in all aspects of my graduate education. I also want to thank my committee members: Dr. Guido Gerig and Dr. Stephen Pizer for their expertise and guidance in image analysis; Dr. Elizabeth Bullitt for her clinical expertise, use of her substantial brain image database, and for funding my studies; and Dr. Andrew Nobel and Dr. J. Steve Marron for their expertise in statistics.

I want to thank all my fellow students (past and present) in the Medical Image Display and Analysis Group (MIDAG) for many helpful discussions, particularly Dr. Michael Rosenthal, Dr. P. Thomas Fletcher, Dr. Sean Ho, Brad Davis, Marcel Prastawa, and Nathan Moon. I would also like to thank MIDAG members Dr. Stephen Aylward and Dr. Mark Foskey for their many constructive comments regarding this work.

I am grateful to the staff of the computer science department for running such a well maintained organization. In particular, I would like to thank Janet Jones for diligently answering my many departmental questions and Delphine Bull whose assistance has saved me many hours. This work was funded in part by the NIBIB-NIH grant R01 EB000219 and the NIH-HLB grant R01 HL69808.

Finally, I would like to thank my family and friends for all their love and support. My mother and stepfather, Marilyn and Chris Albertson, and my father, Willie Lorenzen have tirelessly supported me and my education throughout my entire life. Most important, I would like to thank my soon-to-be born son for the extra incentive to complete this document in a timely manner, and my wife Amy for her unconditional love, support, and encouragement.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

Medical image analysis assists clinicians in many tasks, including the following: evaluating a patient's disease state, evaluating the efficacy of a prescribed treatment, or studying illness through population studies. In their analyses, clinicians draw upon their knowledge of anatomy and their experience viewing medical images and tissues directly. Historically, anatomists have encoded this prior knowledge of anatomy in the form of an *atlas*. Such an atlas may be a set of drawings rendered by experts, such as [75], or a photographic study as in [83]. Digital atlases, the subject of this dissertation, naturally require computers. Several types of digital atlases, including intensity-based, label-based, and probabilistic, are described in [99]. This dissertation focuses on building atlases from multi-modal images.

Brain atlases are often used to study structural neuroanatomy: for surgical planning (e.g., in the planning of tumor resection where statistical atlases provide prior probability models to constrain segmentation to localize critical structures [55]); for instructional purposes [55]; for group studies where atlases are used to study structural differences between populations (e.g., schizophrenics and normal controls [18, 19, 96]); and to study morphological change such as in generative processes (e.g., neonatal brain development [78]) and degenerative processes (e.g., gray matter loss due to Alzheimer's [97, 96]).

The fundamental impetus driving this dissertation is the desire to incorporate as much anatomical information as possible into the problems of medical image registration and atlas formation. Mono-modality images do not capture all the information representing the underlying tissue structure. The information that is provided by a

medical image is dependent on the physics and parameterization of the acquisition process. For example, in computed tomographic imaging the attenuation of x-rays is measured, resulting in excellent delineation of dense bone structure. Magnetic resonance imaging relies on the relaxation properties of the dipole moments of hydrogen atoms under the influence of magnetic fields. The resulting images are, typically, based on three basic tissue parameters (proton density, T1 relaxation time, and T2 relaxation time) as well as flow characteristics. The process involves correlating a series of radio frequency signal measurements with spatial location of the various tissue types. Consequently, magnetic resonance images provide good soft tissue segmentation. The different imaging modalities of magnetic resonance imaging provide complementary information about pathological structures. In the case of certain tumors, T1-weighted images present good fat-tumor contrast whereas T2-weighted images show good tumor-muscle contrast [59]. In multiple scleroris, contrast-enhanced T1-weighted images show the active lesions, T2-weighted images show both new and old lesions, and proton density images highlight lesions near the fluid-filled ventricles [25]. In this dissertation, to utilize the complementary information from multi-modal images, a model-based framework is defined in terms of multi-class posterior probability maps where the classes represent underlying tissue structures.

This dissertation presents a novel framework for multi-class atlas formation from sets of multi-modal images. This framework is extended to the unbiased atlas setting of [49]. These atlases represent anatomical variation present in populations [69, 32, 94]. Many images are mapped into a common coordinate system to study intra-population variability and inter-population differences, to provide voxel-wise mapping of functional sites, and to facilitate tissue and object segmentation via registration of anatomical labels. Atlas formation from a population of medical images is an important problem that is naturally expressed within the subject of *computational anatomy*.

1.2 Computational Anatomy and Atlas Formation

Computational anatomy [32] is the study of anatomical shape variability. Relating anatomical shape to biological growth and function has roots in the seminal work of D'Arcy Thompson in 1917 [93]. Thompson was interested in the comparison of related forms through mathematical transformations rather than in the precise definition of each form. This notion was extended to *deformable templates* [31] in which the space of anatomical imagery is interpreted as the orbit under the group of transformations.

Specifically, computational anatomy belongs to the discipline of geometry governed by pattern theoretic principles whose kinematics, that characterize the transformations, is described in terms of concepts borrowed from continuum mechanics.

In the framework of computational anatomy, the atlas represents the common invariant structure in a population and the transformations relating the coordinate space of the atlas to the coordinate spaces of the individual members of the population encode the variability. This variability represents local structure under deformation or *warping*. An example of this devision into invariance and variability consider the healthy human brain: in every such brain there exist precisely two lateral ventricles with no two pair the same. Therefore, under the computational anatomical framework, an atlas of healthy human brains should represent precisely two lateral ventricles and capture the variability through mathematical transformations that relate them. The atlas formation method proposed in this dissertation falls within this framework.

Understanding anatomical variability requires high-dimensional image registration where the number of parameters used to described the transformations are on the order of the number of spatial elements describing the underlying space of the images. There are many choices in this regard. In terms of ability to capture anatomic variability, transformation models can be loosely dichotomized into small deformation models, which associate energy to distance, and large deformation models, which associate energy to velocity on fluid flows. This dissertation applies the theory of large deformation diffeomorphisms [22, 70, 68] to generate transformations. An important property of diffeomorphic transformations is that they preserve topology. That is, these transformations do not fold or tear space and hence, preserve local structure.

An important component of image registration and, hence, atlas formation, is the choice of image dissimilarity distance function. This is typically a scalar-valued function defined on the Cartesian product $X \times X$ where X is the feature space of the image. The dissimilarity function provides a number indicating how far apart two images are and is chosen based on assumptions about the relationship between the features in the images. When two images are in perfect registration this distance assumes its smallest value. The most natural form of a distance function is a metric.

Definition 1.1 (Metric)**.** *Given a set X, a* metric *on X is function $d : X \times X \to \mathbb{R}$. For all $x, y, z \in X$, this function satisfies the following conditions:*

1. Non-negativity: $d(x, y) \geq 0$
2. Identity of indiscernables: $d(x, y) = 0$ *if and only if* $x = y$.

3. Symmetry: $d(x, y) = d(y, x)$

4. Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$

A simple example of a metric is the Euclidean distance between two points.

This dissertation uses multi-modal images where a statistical relationship is assumed between image features. This suggests the use of information theoretic distances based on Shannon entropy. These distances are typically not true metrics, typically failing property (4) and, often, property (3).

1.3 Thesis and Contributions

> Thesis: *A sharp unbiased multi-tissue class atlas can be constructed from a population of multi-modal image sets, comprised of an arbitrary number of images per subject, using a Bayesian framework and large deformation diffeomorphic registration. Multi-modal image set correspondence can further be achieved via composition of transformations associated with the unbiased atlas.*

The contributions of this dissertation are the following:

1. A theoretical development showing that minimizing sum-of-Kullback-Leibler divergences, in either ordering of parameters, maximizes a lower bound on Bayes probability of error, a measure of indistinguishability between probability distributions.

2. A novel multi-modal image set registration method is presented. To the author's knowledge this is the only method that incorporates an arbitrary number of multi-modal images per subject. An advantageous consequence of this framework is inverse-invariant (symmetric) registration.

3. An extension of the above framework to unbiased multi-class atlas formation[1].

4. The use of information theory to evaluate atlas stability.

5. An application of the atlas formation to an aging study involving multi-modal brain image data from ninety-five subjects.

[1]The work presented in this chapter was done in collaboration with Dr. Sarang Joshi and Brad Davis at the University of North Carolina at Chapel Hill. This work is heavily based on previous papers [63, 49].

1.4 Overview of Chapters

The remainder of this dissertation is organized as follows:

Chapter 2 presents motivates the distances used to drive the multi-modal image set registration and the unbiased multi-class posterior atlas formation. This includes a discussion of Bayes probability of error and the bounds thereof in terms of the sum-of-Kullback-Leibler divergences.

Chapter 3 presents the main theoretical contribution of this dissertation: model-based multi-modal image set registration. The fluid mechanical framework for constraining the resulting transformations to the space of diffeomorphisms is described. Existing information theoretic techniques for image registration will be described with special attention given to multi-variate interpretations of mutual information.

In Chapter 4, using information theoretic distances developed in Chapter 2, the concept of multi-modal image set registration is generalized to unbiased multi-class posterior atlas formation. Existing unbiased atlas formation methods are also discussed.

In Chapter 5, atlas stability, with respect to the number of constituent subjects, is explored.

In Chapter 6, the method from Chapter 4 is applied to MR images from a database of 100 healthy subjects.

Chapter 7 concludes with a discussion of the contributions of this dissertation and possible future work.

Appendix A presents the information theoretic background for Chapter 2.

Appendix B prodives a discussion on the likelihood of the actual optimum in the multi-modal image set registration cost function being achieved.

Chapter 2

Probability Averages

Central to the image registration and atlas formation framework presented in this dissertation is the creation of the *average* or most representative probability from a given collection of probabilities. In this framework, the average probability distribution $\hat{p}$ is defined in terms of minimizing a dispersion measure Γ for a set of distributions $\{p_i\}_{i=1}^N$,

$$\hat{p} = \underset{p}{\operatorname{argmin}} \, \Gamma\left(\{p_i\}_{i=1}^N, p\right). \tag{2.1}$$

This dispersion measure takes the form of a sum of individual distances between p and each p_i,

$$\Gamma\left(\{p_i\}_{i=1}^N, p\right) = \sum_{i=1}^N \gamma\left(\{p_i, p\}\right). \tag{2.2}$$

The distance γ is chosen to be the Kullback-Leibler divergence, as there exists a lower bound on Bayes probability of error P_e between $\hat{p}$ and the individual p_i in terms of Γ. More specifically, minimizing the sum-of-Kullback-Leibler divergences between $\hat{p}$ and each p_i maximizes a lower bound on indistinguishability between the p_i. Since Kullback-Leibler divergence is an asymmetric distance, two sum-of-Kullback-Leibler divergences are considered. The two dispersion measures are

$$\bar{D}\left(\{p_i\}_{i=1}^N || p\right) = \sum_{i=1}^N D(p_i||p) \tag{2.3}$$

and

$$\bar{D}\left(p||\{p_i\}_{i=1}^N\right) = \sum_{i=1}^N D(p||p_i). \tag{2.4}$$

Kullback-Leibler divergence $D(\cdot||\cdot)$ and other fundamental information theoretic measures are defined in Appendix A.

This chapter is organized as follows: Section 2.1 shows that the generalized average

$$\hat{p} = \underset{p}{\operatorname{argmin}}\, \bar{D}\left(\{p_i\}_{i=1}^N||p\right)$$

is the arithmetic mean and that the generalized average

$$\hat{p} = \underset{p}{\operatorname{argmin}}\, \bar{D}\left(p||\{p_i\}_{i=1}^N\right)$$

is the normalized geometric mean; Section 2.2 defines Bayes probability of error; and Section 2.3 presents bounds on Bayes probability of error in terms of the dispersion measures $\bar{D}\left(\{p_i\}_{i=1}^N||\hat{p}\right)$ and $\bar{D}\left(\hat{p}||\{p_i\}_{i=1}^N\right)$ with their respective $\hat{p}$ minimizers.

2.1 Averages from Kullback-Leibler Divergence Sums

This section presents the sum-of-Kullback-Leibler divergence minimizers for the dispersion measures in Equation 2.3 and Equation 2.4. These minimizers have been presented in the literature, for example in [45] for the $N = 2$ case. The derivation for the general case via the method of Langrange multipliers is included for completeness.

2.1.1 Arithmetic Mean from $\bar{D}\left(\{p_i\}_{i=1}^N||p\right)$

Theorem 2.1 ($\bar{D}(\{p_i\}_{i=1}^N||p)$ Minimizer)**.** *Given a set of probability mass functions* $\{p_i\}_{i=1}^N$*, the minimizer*

$$\hat{p} = \underset{p\in\mathcal{P}}{\operatorname{argmin}}\, \bar{D}\left(\{p_i\}_{i=1}^N||p\right),$$

where $\mathcal{P}$ is space of probability mass functions, is the arithmetic mean

$$\hat{p}(x_j) = \frac{1}{N}\sum_{i=1}^{N} p_i(x_j).$$

Proof. This can be shown via the method of Lagrange multipliers. First note that the minimization described can be equivalently expressed as

$$\hat{p} = \underset{p\in\mathcal{P}}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{x\in X} p_i(x)\ln\frac{p_i(x)}{p(x)}$$

subject to constraint $\sum_{x_j\in X}\hat{p}(x_j) = 1$.

Set up the Lagrange multiplier expression to find the minimizer $\hat{p}$

$$L(p,\lambda) = \sum_{i=1}^{N}\sum_{x_j\in X} p_i(x_j)\ln\frac{p_i(x_j)}{p(x_j)} - \lambda\left(\sum_{x_k\in X} p(x_k) - 1\right). \tag{2.5}$$

Solving $\frac{\partial}{\partial p(x_j)}L(p,\lambda) = 0$ for p in terms of λ, yields

$$\begin{aligned}
\frac{\partial}{\partial p(x_j)}L(p,\lambda) &= \sum_{i=1}^{N}\frac{\partial}{\partial p(x_j)}\left[p_i(x_j)\ln p_i(x_j) - p_i(x_j)\ln p(x_j)\right] - \frac{\partial}{\partial p(x_j)}\lambda(p(x_j)-1)\\
&= -\sum_{i=1}^{N}\frac{p_i(x_j)}{p(x_j)} - \lambda\\
\Rightarrow \frac{1}{p(x_j)}\sum_{i=1}^{N} p_i(x_j) &= -\lambda\\
\Rightarrow p(x_j) &= -\frac{1}{\lambda}\sum_{i=1}^{N} p_i(x_j).
\end{aligned} \tag{2.6}$$

As λ is chosen to satisfy the constraint that $p(x)$ is a probability mass function, $-\frac{1}{\lambda} = \frac{1}{N}$. That is,

$$\begin{aligned}
\sum_{x_j\in X} p(x_j) &= 1\\
\Rightarrow \sum_{x_j\in X}\sum_{i=1}^{N} p_i(x_j) &= N.
\end{aligned}$$

Therefore,

$$p(x_j) = \frac{1}{N} \sum_{i=1}^{N} p_i(x_j). \qquad \square$$

With this minimizer the dispersion in Equation 2.3 can be further specified as

$$\begin{aligned} \bar{D}\left(\{p_i\}_{i=1}^N || \hat{p}\right) &= \sum_{i=1}^{N} D\left(p_i || \frac{1}{N} \sum_{j=1}^{N} p_j\right) \\ &= N \cdot JS_\pi(\{p_i\}_{i=1}^N) \end{aligned} \tag{2.7}$$

where JS_π is the generalized Jensen-Shannon divergence with uniform prior $\pi_i = \frac{1}{N}$. Jensen-Shannon divergence is defined in Appendix A.

2.1.2 Normalized Geometric Mean from $\bar{D}\left(p||\{p_i\}_{i=1}^N\right)$

Theorem 2.2 ($\bar{D}(p||\{p_i\}_{i=1}^N)$ Minimizer)**.** *Given a set of probability mass functions* $\{p_i\}_{i=1}^N$*, the minimizer*

$$\hat{p} = \underset{p \in \mathcal{P}}{\operatorname{argmin}} \left(p||\{p_i\}_{i=1}^N\right),$$

where $\mathcal{P}$ *is space of probability mass functions, is the normalized geometric mean*

$$\hat{p}(x_j) = \frac{\left(\prod_{i=1}^N p_i(x_j)\right)^{\frac{1}{N}}}{\sum_{x_k \in X} \left(\prod_{i=1}^N p_i(x_k)\right)^{\frac{1}{N}}}.$$

Proof. This can also be shown via the method of Lagrange multipliers. First note that the minimization described can be equivalently expressed as

$$\hat{p} = \underset{p \in \mathcal{P}}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{x \in X} p(x) \ln \frac{p(x)}{p_i(x)}$$

subject to constraint $\sum_{x_j \in X} \hat{p}(x_j) = 1$. Notice that in this formulation the logarithm used in the definition of $D(\cdot||\cdot)$ has been changed from base two to base e. This reformulation places the minimization problem in a natural setting for variational analysis.

Set up the Lagrange multiplier expression to find minimizer $\hat{p}$

$$\begin{aligned} L(p,\lambda) &= \sum_{i=1}^{N} p \ln \frac{p}{p_i} - \lambda \left(\sum_{x_k \in X} p(x_k) - 1 \right) \\ &= \sum_{i=1}^{N} \sum_{x_j \in X} p(x_j) \ln \frac{p(x_j)}{p_i(x_j)} - \lambda \left(\sum_{x_k \in X} p(x_k) - 1 \right). \end{aligned} \tag{2.8}$$

Solving $\frac{\partial}{\partial p(x_j)} L(p,\lambda) = 0$ for p in terms of λ, yields

$$\begin{aligned} \frac{\partial}{\partial p(x_j)} L(p,\lambda) &= \frac{\partial}{\partial p(x_j)} \left[\sum_{i=1}^{N} \sum_{x_j \in X} p(x_j) \ln \frac{p(x_j)}{p_i(x_j)} - \lambda \left(\sum_{x_k \in X} p(x_k) - 1 \right) \right] \\ &= \sum_{i=1}^{N} \frac{\partial}{\partial p(x_j)} p(x_j) \ln \frac{p(x_j)}{p_i(x_j)} - \frac{\partial}{\partial p(x_j)} \lambda (p(x_j) - 1) \\ &= \sum_{i=1}^{N} \left[\ln \frac{p(x_j)}{p_i(x_j)} + p(x_j) \frac{p_i(x_j)}{p(x_j)} \frac{1}{p_i(x_j)} \right] - \lambda \\ &= \sum_{i=1}^{N} \ln \frac{p(x_j)}{p_i(x_j)} + (N - \lambda) \\ &= \ln \left(\prod_{i=1}^{N} \frac{p(x_j)}{p_i(x_j)} \right) + (N - \lambda) \\ &= 0 \\ \Rightarrow \ln \left(\prod_{i=1}^{N} \frac{p(x_j)}{p_i(x_j)} \right) &= (\lambda - N) \\ \ln \left(\prod_{i=1}^{N} p(x_j) \right) &= (\lambda - N) + \ln \left(\prod_{i=1}^{N} p_i(x_j) \right) \\ \Rightarrow p(x_j)^N &= e^{(\lambda - N)} \prod_{i=1}^{N} p_i(x_j) \\ \Rightarrow p(x_j) &= e^{\left(\frac{\lambda}{N} - 1\right)} \left(\prod_{i=1}^{N} p_i(x_j) \right)^{\frac{1}{N}}. \end{aligned} \tag{2.9}$$

Continuing with the Lagrange multiplier method, next let

$$\Gamma_j = \left(\prod_{i=1}^{N} p_i(x_j) \right)^{\frac{1}{N}}$$

and, hence, $p(x_j) = e^{\left(\frac{\lambda}{N}-1\right)}\Gamma_j$. Substituting this expression back into Equation 2.8, yields

$$\begin{aligned}
L(p,\lambda) &= \sum_{i=1}^{N} \sum_{x_j \in X} e^{\left(\frac{\lambda}{N}-1\right)} \Gamma_j \ln \frac{e^{\left(\frac{\lambda}{N}-1\right)}\Gamma_j}{p_i(x_j)} - \lambda \left(\sum_{x_k \in X} e^{\left(\frac{\lambda}{N}-1\right)}\Gamma_k - 1 \right) \\
&= e^{\left(\frac{\lambda}{N}-1\right)} \sum_{i=1}^{N} \sum_{x_j \in X} \Gamma_j \ln e^{\left(\frac{\lambda}{N}-1\right)} + e^{\left(\frac{\lambda}{N}-1\right)} \sum_{i=1}^{N} \sum_{x_j \in X} \Gamma_j \ln \Gamma_j \\
&\quad - e^{\left(\frac{\lambda}{N}-1\right)} \sum_{i=1}^{N} \sum_{x_j \in X} \Gamma_j \ln p_i(x_j) - \lambda \sum_{x_k \in X} e^{\left(\frac{\lambda}{N}-1\right)}\Gamma_k + \lambda \\
&= \left(\frac{\lambda}{N} - 1 \right) e^{\left(\frac{\lambda}{N}-1\right)} N \sum_{x_j \in X} \Gamma_j + N e^{\left(\frac{\lambda}{N}-1\right)} \sum_{x_j \in X} \Gamma_j \ln \Gamma_j \\
&\quad - e^{\left(\frac{\lambda}{N}-1\right)} \sum_{x_j \in X} \Gamma_j \ln \left(\prod_{i=1}^{N} p_i(x_j) \right) - \lambda e^{\left(\frac{\lambda}{N}-1\right)} \sum_{x_k \in X} \Gamma_k + \lambda.
\end{aligned}$$

Note that since

$$\begin{aligned}
\sum_{x_j \in X} \Gamma_j \ln \left(\prod_{i=1}^{N} p_i(x_j) \right) &= \sum_{x_j \in X} \Gamma_j \ln \Gamma_j^N \\
&= N \sum_{x_j \in X} \Gamma_j \ln \Gamma_j
\end{aligned}$$

the above equation simplifies to

$$\begin{aligned}
L(p,\lambda) &= \left[\left(\frac{\lambda}{N} - 1 \right) N - \lambda \right] e^{\left(\frac{\lambda}{N}-1\right)} \sum_{x_j \in X} \Gamma_j + \lambda \\
&= -N e^{\left(\frac{\lambda}{N}-1\right)} \sum_{x_j \in X} \Gamma_j + \lambda \\
&= -N e^{\left(\frac{\lambda}{N}-1\right)} \sum_{x_j \in X} \left(\prod_{i=1}^{N} p_i(x_j) \right)^{\frac{1}{N}} + \lambda
\end{aligned}$$

Taking the variation with respect to λ, note that

$$\begin{aligned}\frac{\partial}{\partial\lambda}L(p,\lambda) &= \frac{\partial}{\partial\lambda}\left[-Ne^{\left(\frac{\lambda}{N}-1\right)}\sum_{x_j\in X}\left(\prod_{i=1}^{N}p_i(x_j)\right)^{\frac{1}{N}}+\lambda\right] \\ &= -e^{\left(\frac{\lambda}{N}-1\right)}\sum_{x_j\in X}\left(\prod_{i=1}^{N}p_i(x_j)\right)^{\frac{1}{N}}+1.\end{aligned}$$

Setting $\frac{\partial}{\partial\lambda}L(p,\lambda)=0$ implies

$$\begin{aligned}e^{\left(\frac{\lambda}{N}-1\right)}\sum_{x_j\in X}\left(\prod_{i=1}^{N}p_i(x_j)\right)^{\frac{1}{N}} &= 1 \\ \Rightarrow e^{\left(\frac{\lambda}{N}-1\right)} &= \frac{1}{\sum_{x_j\in X}\left(\prod_{i=1}^{N}p_i(x_j)\right)^{\frac{1}{N}}}.\end{aligned}$$

Finally, substituting this expression for $e^{\left(\frac{\lambda}{N}-1\right)}$ back into Equation 2.9 results in

$$\begin{aligned}p(x_j) &= e^{\left(\frac{\lambda}{N}-1\right)}\left(\prod_{i=1}^{N}p_i(x_j)\right)^{\frac{1}{N}} \\ &= \frac{\left(\prod_{i=1}^{N}p_i(x_j)\right)^{\frac{1}{N}}}{\sum_{x_k\in X}\left(\prod_{i=1}^{N}p_i(x_k)\right)^{\frac{1}{N}}}.\end{aligned}$$

□

With this minimizer the dispersion in Equation 2.4 can be further specified as

$$
\begin{aligned}
\bar{D}\left(\hat{p}||\{p_i\}_{i=1}^N\right) &= \sum_{i=1}^N D(\hat{p}||p_i) \\
&= \sum_{i=1}^N \sum_{x_j \in X} \hat{p}(x_j) \log \frac{\hat{p}(x_j)}{p_i(x_j)} \\
&= \sum_{x_j \in X} \hat{p}(x_j) \sum_{i=1}^N \log \frac{\hat{p}(x_j)}{p_i(x_j)} \\
&= \sum_{x_j \in X} \hat{p}(x_j) \log \left(\prod_{i=1}^N \frac{\hat{p}(x_j)}{p_i(x_j)} \right) \\
&= \sum_{x_j \in X} \hat{p}(x_j) \log \left(\hat{p}(x_j)^N \prod_{i=1}^N \frac{1}{p_i(x_j)} \right) \\
&= \sum_{x_j \in X} \hat{p}(x_j) \log \left(\frac{\prod_{i=1}^N p_i(x_j)}{\left(\sum_{x_k \in X} \left(\prod_{i=1}^N p_i(x_k) \right)^{\frac{1}{N}} \right)^N} \prod_{i=1}^N \frac{1}{p_i(x_j)} \right) \\
&= -N \sum_{x_j \in X} \hat{p}(x_j) \log \sum_{x_k \in X} \left(\prod_{i=1}^N p_i(x_k) \right)^{\frac{1}{N}} \\
&= \left(-N \log \sum_{x_k \in X} \left(\prod_{i=1}^N p_i(x_k) \right)^{\frac{1}{N}} \right) \left(\sum_{x_j \in X} p(x_j) \right) \\
&= -N \log \sum_{x_k \in X} \left(\prod_{i=1}^N p_i(x_k) \right)^{\frac{1}{N}} . \qquad (2.10)
\end{aligned}
$$

2.2 Bayes Probability of Error

This section defines Bayes probability of error in terms of an N-hypothesis decision-theory problem. In the probability average estimation setting, the hypotheses are the individual probabilities of a population and their probability average. In the Section 2.3, bounds on Bayes probability of error in terms of the dispersion measures $\bar{D}\left(\{p_i\}_{i=1}^N||\hat{p}\right)$ and $\bar{D}\left(\hat{p}||\{p_i\}_{i=1}^N\right)$ with their respective $\hat{p}$ minimizers is presented.

Consider the N-hypothesis decision-theory problem of classifying an observation as coming from one of N possible hypotheses $\{H_i\}_{i=1}^N$. Let $\{\pi_i\}_{i=1}^N$ denote the prior

probabilities associated with these N hypotheses, and let $\{p(x|H_i)\}_{i=1}^N$ denote the event conditional probability distributions given the hypotheses. Using Bayes' Rule, for an observed event, x, the posterior probability of H_i is

$$\begin{aligned} p(H_i|x) &= \frac{p(x|H_i)\pi_i}{p(x)} \\ &= \frac{p(x|H_i)\pi_i}{\sum_{k=1}^N p(x|H_k)\pi_k}. \end{aligned}$$

To minimize the probability of selecting the incorrect hypothesis, choose the hypothesis with the largest posterior probability. This has the associated conditional probability of error

$$P_e(x|H_i) = 1 - \max_i\{p(H_i|x)\}.$$

With this notion of conditional probability of error the Bayes probability of error (or indistinguishability) can be defined.

Definition 2.1 (Bayes Probability of Error)**.** Bayes probability of error *is the expected conditional probability of error*

$$\begin{aligned} P_e\left(\{p_i, \pi_i\}_{i=1}^N\right) &= \mathcal{E}_p\left[P_e(x|H_i)\right] \\ &= \sum_{x\in X} p(x)\left(1 - \max_i\{p(H_i|x)\}\right) \\ &= \sum_{x\in X} p(x) - \sum_{x\in X} p(x)\max_i\left\{\frac{p(x|H_i)\pi_i}{p(x)}\right\} \\ &= 1 - \sum_{x\in X}\max_i\{p(x|H_i)\pi_i\}. \end{aligned}$$

This error results when one has complete knowledge of the probability distributions with which to construct the optimal decision rule, that is, to select the hypothesis H_i for which the posterior $p(H_i|x)$ is maximal. Figure 2.1 graphically shows Bayes probability of error for a set of four distributions. In practice, real distributions are not known so P_e cannot be directly computed. Therefore bounds on P_e are desired.

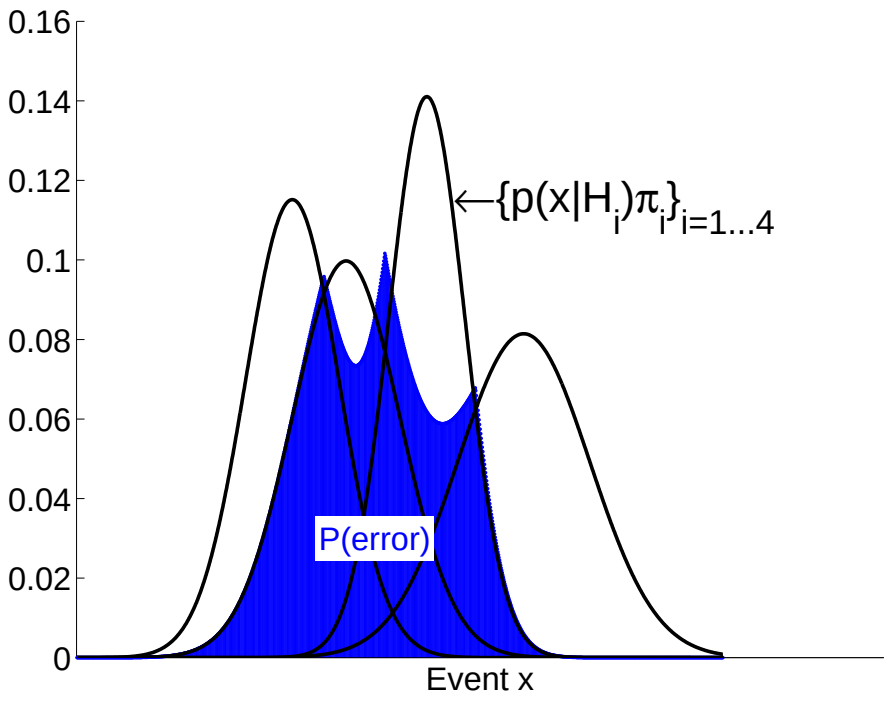


Figure 2.1: Bayes Probability of Error Example

Let $\{\pi_i\}_{i=1...4}$ be the a priori probabilities of hypotheses $\{H_i\}_{i=1...4}$ being true and $\{p(x|H_i)\}_{i=1...4}$ be the hypothesis-conditional likelihoods for event x. In this example, the distributions are continuous and the prior probabilities are taken to be equal. The blue shaded area represents the Bayes probability of misclassification error. In the context of neuroanatomical matching, H_i could represent individual subjects, and x could represent tissue structures such as grey matter, white matter, and cerebrospinal fluid.

Under the uniform prior $\pi_i = \frac{1}{N}$ condition the Bayes probability of error becomes

$$\begin{aligned} P_e\left(\{p_i, \pi_i\}_{i=1}^N\right) &= 1 - \sum_{x \in X} \max_i \{p(x|H_i)\frac{1}{N}\} \\ &= 1 - \frac{1}{N}\sum_{x \in X} \max_i \{p(x|H_i)\}. \end{aligned}$$

If the $p(x|H_i)$ are identical, then the probability of error assumes its greatest value at $P_e = \frac{N-1}{N}$. For example, if $N = 4$, then the probability of selecting $p(x|H_j)$ from the set of identical probabilities $\{p(x|H_i)\}_{i=1\ldots4}$ is $P_e = \frac{3}{4}$. In this case, as $N \to \infty$, $P_e \to 1$.

2.3 Bounds on P_e

2.3.1 $\bar{D}\left(\{p_i\}_{i=1}^N || \hat{p}\right)$ Bounds on P_e

Theorem 2.3 (Jensen Shannon P_e Bounds)**.** *Given a set of probability mass functions* $\{p_i\}_{i=1}^N$ *and associated priors* $\{\pi_i\}_{i=1}^N$ *with* $\sum_{i=1}^N \pi_i = 1$*,*

$$\frac{1}{4(N-1)} J\left(\{p_i, \pi_i\}_{i=1}^N\right)^2 \leq P_e\left(\{p_i, \pi_i\}_{i=1}^N\right) \leq \frac{1}{2} J\left(\{p_i, \pi_i\}_{i=1}^N\right)$$

where $J\left(\{p_i, \pi_i\}_{i=1}^N\right) = H(\pi) - JS_\pi\left(\{p_i\}_{i=1}^N\right)$ *and* $H(\pi) = -\sum_{i=1}^N \pi_i \log \pi_i$*.*

Proof. See [60] for an argument involving conditional entropy bounds on P_e [41]. □

It immediately follows from Equation 2.7 and Theorem 2.3 that

$$\frac{1}{4(N-1)}\left(H(\pi) - \frac{1}{N}\bar{D}\left(\{p_i\}_{i=1}^N || \hat{p}\right)\right)^2 \leq P_e\left(\{p_i, \pi_i\}_{i=1}^N\right) \leq \frac{1}{2}\left(H(\pi) - \frac{1}{N}\bar{D}\left(\{p_i\}_{i=1}^N || \hat{p}\right)\right)$$

where

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^N p_i(x).$$

Therefore, minimizing the sum-of-Kullback-Leibler divergence, $\bar{D}\left(\{p_i\}_{i=1}^N||\hat{p}\right)$, maximizes $\left(H(\pi) - \frac{1}{N}\bar{D}\left(\{p_i\}_{i=1}^N||\hat{p}\right)\right)^2$, and, hence, maximizes the lower bound on P_e, or indistinguishability between the p_i.

Under the uniform prior $\pi_i = \frac{1}{N}$ condition with $\bar{D}\left(\{p_i\}_{i=1}^N||\hat{p}\right) = 0$ the lower bound becomes

$$\frac{1}{4(N-1)}\left(H(\pi) - \frac{1}{N}\bar{D}\left(\{p_i\}_{i=1}^N||\hat{p}\right)\right)^2 = \frac{(\log N)^2}{4(N-1)}.$$

In this case,

$$\begin{aligned}\lim_{N\to\infty}\frac{(\log N)^2}{4(N-1)} &= \lim_{N\to\infty}\frac{\log N}{2N}\\ &= \lim_{N\to\infty}\frac{1}{2N}\\ &= 0.\end{aligned}$$

Although, for a given N, minimizing $\bar{D}\left(\{p_i\}_{i=1}^N||\hat{p}\right)$ maximizes a lower bound on P_e, as $N \to \infty$, that lower bound $\frac{(\log N)^2}{4(N-1)} \to 0$. A much tighter, and more meaningful, bound on P_e can be defined using $\bar{D}\left(\hat{p}||\{p_i\}_{i=1}^N\right)$

2.3.2 $\bar{D}\left(\hat{p}||\{p_i\}_{i=1}^N\right)$ Bounds on $\bar{P}_e$

To find a lower bound on Bayes probability of error based on $\bar{D}\left(\hat{p}||\{p_i\}_{i=1}^N\right)$ it is convenient to consider the average of pair-wise probability of errors between individual p_i. That is, a bound on

$$\bar{P}_e\left(\{p_i\}_{i=1}^N\right) = \frac{1}{N}\sum_{i=1}^{N} P_e(\{p_i, p_{(i \bmod N)+1}\})$$

is desired where $P_e(\{p_i, p_j\})$ is the Bayes probability of error when only p_i and p_j are involved. From equation 2.10, note that

$$\begin{aligned}\bar{D}\left(\hat{p}||\{p_i, p_j\}\right) &= D(\hat{p}, p_i) + D(\hat{p}, p_j)\\ &= -2\log\sum_{x\in X}\sqrt{p_i(x)p_j(x)}\end{aligned}$$

where

$$\hat{p}(x) = \frac{\sqrt{p_i(x)p_j(x)}}{\sum_{x' \in X} \sqrt{p_i(x')p_j(x')}},$$

the normalized geometric mean of p_i and p_j.

Define the sum of pair-wise distances

$$\begin{aligned} D^* \left(\{p_i\}_{i=1}^N \right) &= \sum_{i=1}^{N} \bar{D}(\hat{p} || \{p_i, p_{(i \bmod N)+1}\}) \\ &= -2 \sum_{i=1}^{N} \log \sum_{x \in X} \sqrt{p_i(x) p_{i(\bmod)+1}(x)}. \end{aligned}$$

The next section will show that $D^* \left(\{p_i\}_{i=1}^N \right)$ is bounded above by $\bar{D} \left(\hat{p} || \{p_i\}_{i=1}^N \right)$. The implication is that any optimization problem that minimizes $\bar{D} \left(\hat{p} || \{p_i\}_{i=1}^N \right)$ also minimizes $D^* \left(\{p_i\}_{i=1}^N \right)$. Additionally, it will be shown that $D^* \left(\{p_i\}_{i=1}^N \right)$ provides bounds on $\bar{P}_e$.

2.3.2.1 $D^* \left(\{p_i\}_{i=1}^N \right) \leq \bar{D} \left(\hat{p} || \{p_i\}_{i=1}^N \right)$

Theorem 2.4 (Generalized Hölder's Inequality)**.** *Let $r_i \geq 1$ such that $\sum_{i=1}^N \frac{1}{r_i} = 1$. Then the following inequality holds for $f_i(x) \geq 0$,*

$$\sum_{x \in X} \prod_{i=1}^{N} f_i(x)^{\frac{1}{r_i}} \leq \prod_{i=1}^{N} \left(\sum_{x \in X} f_i(x) \right)^{\frac{1}{r_i}}.$$

Proof. For an argument using generalized arithmetic means see [37]. □

Proposition 2.1. *Let $\{p_i(x)\}_{i=1}^N$ be a set of probability mass functions defined on random variable X. Then*

$$\sum_{x \in X} \prod_{k=1}^{N} p_k(x)^{\frac{1}{N}} \leq \left(\sum_{x \in X} \sqrt{p_i(x)p_j(x)} \right)^{\frac{2}{N}}$$

for any $i, j \in \{1, \ldots, N\}$ where $i \neq j$.

Proof. By applying Theorem 2.4,

$$\sum_{x\in X}\prod_{k=1}^{N} p_k(x)^{\frac{1}{N}} = \sum_{x\in X}\left(\sqrt{p_i(x)p_j(x)}\right)^{\frac{2}{N}} \prod_{\substack{k=1\\k\neq i,j}}^{N} p_k(x)^{\frac{1}{N}}$$

$$\leq \left(\sum_{x\in X}\sqrt{p_i(x)p_j(x)}\right)^{\frac{2}{N}} \prod_{\substack{k=1\\k\neq i,j}}^{N} \left(\sum_{x\in X} p_k(x)\right)^{\frac{1}{N}}$$

Using the fact that each $p_k(x)$ sum to one,

$$\sum_{x\in X}\prod_{k=1}^{N} p_k(x)^{\frac{1}{N}} = \left(\sum_{x\in X}\sqrt{p_i(x)p_j(x)}\right)^{\frac{2}{N}}. \qquad \square$$

Using the inequality in Proposition 2.1, one can now show that $D^*\left(\{p_i\}_{i=1}^N\right) \leq \bar{D}\left(\hat{p}||\{p_i\}_{i=1}^N\right)$.

Theorem 2.5 ($\bar{D}_(\hat{p}||\{p_i\}_{i=1}^N$) Lower Bound)**.** *Let $\{p_i(x)\}_{i=1}^N$ be a set of probability mass functions defined on random variable X. Then*

$$D^*\left(\{p_i\}_{i=1}^N\right) \leq \bar{D}\left(\hat{p}||\{p_i\}_{i=1}^N\right).$$

Proof. This immediately follows from Proposition 2.1 and the definitions of $D^*\left(\{p_i\}_{i=1}^N\right)$ and $\bar{D}\left(\hat{p}||\{p_i\}_{i=1}^N\right)$,

$$\sum_{x\in X}\prod_{k=1}^{N} p_k(x)^{\frac{1}{N}} \leq \left(\sum_{x\in X}\sqrt{p_i(x)p_j(x)}\right)^{\frac{2}{N}}$$

$$\Rightarrow \sum_{x\in X}\prod_{k=1}^{N} p_k(x)^{\frac{1}{N}} \leq \left(\sum_{x\in X}\sqrt{p_i(x)p_{(i \bmod N)+1}(x)}\right)^{\frac{2}{N}}$$

Taking the logarithm of both sides of the above inequality and simplifying yields

$$N \log \sum_{x \in X} \prod_{k=1}^{N} p_k(x)^{\frac{1}{N}} \leq 2 \sum_{i=1}^{N} \log \sum_{x \in X} \sqrt{p_i(x) p_{(i \bmod N)+1}(x)}$$

$$\Rightarrow -2 \sum_{i=1}^{N} \log \sum_{x \in X} \sqrt{p_i(x) p_{(i \bmod N)+1}(x)} \leq -N \log \sum_{x \in X} \prod_{k=1}^{N} p_k(x)^{\frac{1}{N}}$$

$$\Rightarrow \sum_{i=1}^{N} \bar{D}(\hat{p} || p_i, p_{(i \bmod N)+1}) \leq \bar{D}\left(\hat{p} || \{p_i\}_{i=1}^{N}\right). \qquad \square$$

2.3.2.2 $\bar{D}(\hat{p}||p_i, p_{(i \bmod N)+1})$ Bounds on $P_e(\{p_i, p_j\})$

Definition 2.2 (Bhattacharyya Coefficient)**.** *The Bhattacharyya coefficient for two densities $p(x)$ and $q(x)$ is defined by*

$$\rho(p, q) = \sum_{x \in X} \sqrt{p(x) q(x)}.$$

The Bhattacharyya coefficient is a divergence-type measure which can be geometrically interpreted as the cosine of the angle between n-dimensional vectors.

Theorem 2.6 (Bhattacharrya Bounds on $P(error)$)**.** *Let p_1 and p_2 be two probability distributions and $\rho = \rho(p_1, p_2)$ be the Bhattacharrya coefficient defined by them. Then the Bayes probability of error for p_1 and p_2 is bounded as follows:*

$$\frac{1}{2}\left(1 - \sqrt{1 - \rho^2}\right) \leq P_e\left(\{p_1, p_2\}\right) \leq \frac{1}{2}\rho.$$

Proof. See [53] for an argument involving Kolmogorov variational distance. $\square$

The $\bar{D}(\hat{p}||\{p_i, p_j\})$ bounds on $P_e\left(\{p_i, p_j\}\right)$ can now be computed as

$$\bar{D}(\hat{p}||\{p_i, p_j\}) = -2 \log \rho$$

$$\Rightarrow e^{-\frac{1}{2}\bar{D}(\hat{p}||\{p_i, p_j\})} = \rho$$

and, by Theorem 2.6,

$$\frac{1}{2}\left(1 - \sqrt{1 - e^{-\bar{D}(\hat{p}||\{p_i, p_j\})}}\right) \leq P_e\left(\{p_i, p_j\}\right) \leq \frac{1}{2} e^{-\frac{1}{2}\bar{D}(\hat{p}||\{p_i, p_j\})}.$$

When $\bar{D}(\hat{p}||\{p_i, p_j\}) = 0$, $P_e(\{p_i, p_j\}) = \frac{1}{2}$ which is what is expected, namely $p_i(x) = p_j(x)$ for all x.

2.3.2.3 $\bar{D}\left(\hat{p}||\{p_i\}_{i=1}^N\right)$ Bounds on $\bar{P}_e$

The following inequality is useful in providing bounds on $\bar{P}_e$ based on $\bar{D}\left(\hat{p}||\{p_i\}_{i=1}^N\right)$.

Proposition 2.2 (Bounds on e^x)**.** *For $x \geq 0$,*

$$1 - x \leq e^{-x} \leq 1 - x + \frac{1}{2}x^2.$$

Proof. The proof follows from considering the Taylor expansion of e^{-x}. □

With Proposition 2.2, the following theorem can now be proven.

Theorem 2.7 ($\bar{D}(\hat{p}||\{p_i\}_{i=1}^N)$ Bounds on $\bar{P}_e$)**.** *Let $\{p_i(x)\}_{i=1}^N$ be a set of probability mass functions on random variable X. Then*

$$\frac{1}{2} - \frac{1}{2N}\sqrt{\bar{D}\left(\hat{p}||\{p_i\}_{i=1}^N\right)} \leq \bar{P}_e\left(\{p_i\}_{i=1}^N\right) \leq \frac{1}{2} + \frac{1}{16N}\bar{D}\left(\hat{p}||\{p_i\}_{i=1}^N\right)^2$$

where

$$\bar{P}_e\left(\{p_i\}_{i=1}^N\right) = \frac{1}{N}\sum_{i=1}^{N} P_e(\{p_i, p_{(i \bmod N)+1}\}),$$

with $P_e(\{p_i, p_j\})$ the Bayes probability of error between p_i and p_j, and

$$\bar{D}\left(\hat{p}||\{p_i\}_{i=1}^N\right) = -N \log \sum_{x_k \in X} \left(\prod_{i=1}^{N} p_i(x_k)\right)^{\frac{1}{N}}.$$

Proof. Lower Bounds:

$$\frac{1}{2}\left(1-\sqrt{1-e^{-\bar{D}(\hat{p}||\{p_i,p_j\})}}\right) \leq P_e\left(\{p_i,p_j\}\right)$$

$$\Rightarrow \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left(1-\sqrt{1-e^{-\bar{D}(\hat{p}||\{p_i,p_{(i \bmod N)+1}\})}}\right) \leq \bar{P}_e\left(\{p_i\}_{i=1}^{N}\right)$$

$$\Rightarrow \frac{1}{2}-\frac{1}{2N}\sum_{i=1}^{N}\sqrt{1-e^{-\bar{D}(\hat{p}||\{p_i,p_{(i \bmod N)+1}\})}} \leq \bar{P}_e\left(\{p_i\}_{i=1}^{N}\right)$$

$$\Rightarrow \frac{1}{2}-\frac{1}{2N}\sqrt{\sum_{i=1}^{N}\left(1-e^{-\bar{D}(\hat{p}||\{p_i,p_{(i \bmod N)+1}\})}\right)} \leq \bar{P}_e\left(\{p_i\}_{i=1}^{N}\right)$$

$$\Rightarrow \frac{1}{2}-\frac{1}{2N}\sqrt{N-\sum_{i=1}^{N}e^{-\bar{D}(\hat{p}||\{p_i,p_{(i \bmod N)+1}\})}} \leq \bar{P}_e\left(\{p_i\}_{i=1}^{N}\right).$$

Using the second inequality of Proposition 2.2, note that

$$\frac{1}{2}-\frac{1}{2N}\sqrt{N-\sum_{i=1}^{N}(1-\bar{D}(\hat{p}||\{p_i,p_{(i \bmod N)+1}\}))} \leq \bar{P}_e\left(\{p_i\}_{i=1}^{N}\right)$$

$$\Rightarrow \frac{1}{2}-\frac{1}{2N}\sqrt{D^*\left(\{p_i\}_{i=1}^{N}\right)} \leq \bar{P}_e\left(\{p_i\}_{i=1}^{N}\right)$$

$$\Rightarrow \frac{1}{2}-\frac{1}{2N}\sqrt{\bar{D}\left(\hat{p}||\{p_i\}_{i=1}^{N}\right)} \leq \bar{P}_e\left(\{p_i\}_{i=1}^{N}\right)$$

since $\bar{D}\left(\hat{p}||\{p_i\}_{i=1}^{N}\right) \geq D^*\left(\{p_i\}_{i=1}^{N}\right)$.

The first inequality of Proposition 2.2 yields an upper bound on $\bar{P}_e\left(\{p_i\}_{i=1}^{N}\right)$,

$$
\begin{aligned}
P_e\left(\{p_i, p_j\}\right) &\leq \frac{1}{2} e^{-\frac{1}{2}\bar{D}(\hat{p}||\{p_i,p_j\})} \\
\Rightarrow \bar{P}_e\left(\{p_i\}_{i=1}^N\right) &\leq \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2} e^{-\frac{1}{2}\bar{D}(\hat{p}||\{p_i,p_{(i \bmod N)+1}\})} \\
&= \frac{1}{2N}\sum_{i=1}^{N} e^{-\frac{1}{2}\bar{D}(\hat{p}||\{p_i,p_{(i \bmod N)+1}\})} \\
&\leq \frac{1}{2N}\sum_{i=1}^{N} \left(1 - \frac{1}{2}\bar{D}(\hat{p}||\{p_i, p_{(i \bmod N)+1}\}) + \frac{1}{2}\left(\frac{1}{2}\bar{D}(\hat{p}||\{p_i, p_{(i \bmod N)+1}\})\right)^2\right) \\
&= \frac{1}{2} - \frac{1}{4N} D^*\left(\{p_i\}_{i=1}^N\right) + \frac{1}{16N}\sum_{i=1}^{N} \bar{D}(\hat{p}||\{p_i, p_{(i \bmod N)+1}\})^2 \\
&\leq \frac{1}{2} - \frac{1}{4N} D^*\left(\{p_i\}_{i=1}^N\right) + \frac{1}{16N}\left(\sum_{i=1}^{N} \bar{D}(\hat{p}||\{p_i, p_{(i \bmod N)+1}\})\right)^2 \\
&= \frac{1}{2} - \frac{1}{4N} D^*\left(\{p_i\}_{i=1}^N\right) + \frac{1}{16N} D^*\left(\{p_i\}_{i=1}^N\right)^2 \\
&\leq \frac{1}{2} + \frac{1}{16N} D^*\left(\{p_i\}_{i=1}^N\right)^2 \\
&\leq \frac{1}{2} + \frac{1}{16N} \bar{D}\left(\hat{p}||\{p_i\}_{i=1}^N\right)^2
\end{aligned}
$$

since $D^*\left(\{p_i\}_{i=1}^N\right) \leq \bar{D}\left(\hat{p}||\{p_i\}_{i=1}^N\right)$. □

Therefore, minimizing $\bar{D}\left(\hat{p}||\{p_i\}_{i=1}^N\right)$ maximizes the lower bound on $\bar{P}_e\left(\{p_i\}_{i=1}^N\right)$, which, in turn, maximizes the Bhattacharrya lower bound on P_e in Theorem 2.6. When $\bar{D}\left(\hat{p}||\{p_i\}_{i=1}^N\right) = 0$ the lower bound becomes $\frac{1}{2}$, which is the Bayes probability of error between two equally weighted and identical probabilities. Recall that $\bar{P}_e$ is an average of N pair-wise Bayes probability of errors so this lower bound is tight.

Chapter 3

Multi-Modal Image Set Registration[1]

In this chapter the main methodological contribution of this dissertation is presented: a framework for the registration of multi-modal image sets. The chapter begins by defining multi-modal image sets and motivating their use. The registration method is driven by the sum-of-Kullback-Leibler divergences, $\bar{D}\left(\hat{p}||\{p_i\}_{i=1}^N\right)$, introduced as Definition 2.10 in Chapter 2. Although $\bar{D}\left(\hat{p}||\{p_i\}_{i=1}^N\right)$ was defined for the general N-observation setting, for clarity first consider the $N = 2$ case for registration between two sets of multi-modal images. In Chapter 4, this framework is extended to general $N-$way registration for atlas formation.

Modern imaging techniques provide an array of imaging modalities that enable the acquisition of complementary information representing an underlying anatomy. To utilize this information, numerous multi-modal image registration algorithms have been developed. Most of these algorithms find a mapping between two scalar images. To utilize multiple scalar multi-modal images of a single anatomy, however, define a multi-modal image set, $\bar{I}$, as a collection of m co-registered multi-modal images where, for a given spatial location x, $\bar{I}(x) \in \mathbb{R}^m$. Figure 3.1 is a cartoon depicting two image sets each with a different number of constituent scalar multi-modal images. Throughout this dissertation, it is assumed that, for a given subject, the multi-modal images of that subject are co-registered.

The problem of multi-modal image set registration is defined as follows: find a mapping between two subjects for each of which is defined a tuple of multi-modal scalar-valued images, $\bar{I}_1$ and $\bar{I}_2$. More specifically, find a mapping that best matches structure, subject to certain penalties, typically smoothness of the transformations. Mathematically, find the mappings $f : \Omega_1 \rightarrow \Omega_2$ and $g : \Omega_2 \rightarrow \Omega_1$ where Ω_1 and Ω_2

[1]This chapter is an extension of portions of the recent MedIA paper on the topic [63].

are the domains of image sets $\bar{I}_1$ and $\bar{I}_2$ respectively. This arrangement is depicted in Figure 3.2. For the registration to be symmetric or inverse-invariant, the transformation compositions $f \circ g$ and $g \circ f$ must result in the identity map.

The multi-modal image set registration has potential significance in various applications in medical image analysis that rely on the measurement of image sets. Multi-modal MR imaging is standard in the protocol for evaluating pathologies such as tumors and lesions. In this dissertation, results from T1-weighted, T2-weighted, and MR angiography are presented. Other, intrinsically multi-modal modal environments such as combined inter-operative CT/PET acquisitions would provide such a setting for multi-modal image set registration. Registration between images of individuals presenting pathology and images of healthy subjects is a challenging task since space-occupying lesions have to be treated differently from infiltrating lesions. Specifically, the registration needs to accommodate both local spatial deformation and local change of image intensity. Existing registration methods involving scalar images based on image brightness do not accommodate pathologies. Another potential application for image set registration is the registration of images acquired from scanners of different field strength. Image set registration across different scanners becomes an increasingly important component in multi-center studies which investigate developmental changes covering multiple years and follow-up studies of diseases with change of scanner technology. Images acquired from different scanners potentially have different contrasts and different spatial distortions. In functional studies, it is often desirable to register functional data with structural data. Typically, a time series average is registered to an anatomical image of the same subject. However, in producing the average functional image, some information is lost. Therefore, a multivariate approach is desired [2]. The registration method presented in this dissertation may address these problems as the registration is based on underlying anatomical structure (e.g., tissue) rather than on image intensities. These anatomical structures are modeled as class-conditional probability maps where, at each spatial location, a class-conditional probability mass function is defined. In the formation of the class probabilities, classes can be explicitly assigned to the various healthy and pathological tissues. This allows us to potentially model the behaviors of the different tissues during the registration process.

Before discussing the problem of multi-modal image set registration, common methods for multi-modal scalar image registration in Section 3.1 are reviewed. Particular attention is given to image registration based on mutual information in Section 3.1.1. Multi-modal image set registration and the implications of using multivariate mutual

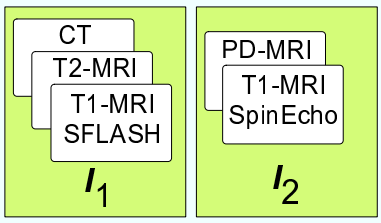


Figure 3.1: Two Sample Image Sets

An image set is a collection of co-registered multi-modal scalar images.

information are developed in Section 3.2. Existing multiple image registration methods and joint segmentation and registration methods are presented in Sections 3.2.2 and 3.2.3 respectively. In Section 3.2.4, the multi-modal image set registration framework is described. Finally, in Section 3.3, results from several 3D multi-modal image set registration experiments are presented. Next follows a brief overview of information theoretic image registration algorithms.

3.1 Multi-Modal Image Registration Background

Many image registration techniques for multi-modal images involve information theoretic distance measures on the space of probability distribution functions. Consequently, a probability mass function is the typical feature to be matched during registration. Mutual information is one method that is typically used to register multi-modal images and will be discussed in Section 3.1.1.

A method that minimizes Kullback-Leibler divergence between expected and observed joint class histograms is presented in [13]. A joint class histogram between the image pairs is estimated by assigning each bin value equal to the total number of occurrences of the corresponding class label pairs. This technique, however, estimates class labels as a preprocessing step and is used only for rigid registration between scalar images. Another class-label approach uses joint voxel class distribution matrices that represent local fuzzy correspondence of object class labels for pairs of corresponding voxels is presented in [20]. This method also uses Kullback-Leibler divergence as a dissimilarity criterion and is presenting in a non-linear fluid registration framework with Gaussian regularization of the estimated transformations. The method presented in

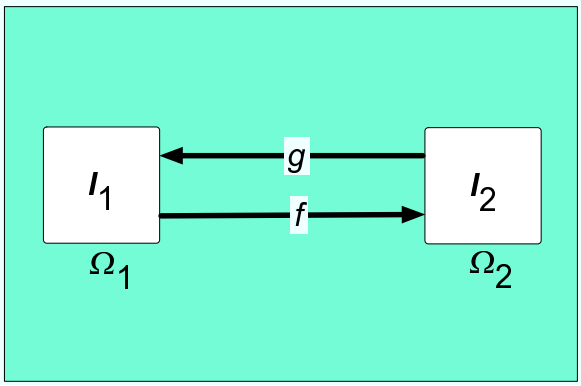


Figure 3.2: Image Set Registration

Registration of image sets $\bar{I}_1$ and $\bar{I}_2$: finding a forward transformation f that maps the domain of $\bar{I}_1$ to the domain of $\bar{I}_2$ and a reverse transformation g that maps the domain of $\bar{I}_2$ to the domain of $\bar{I}_1$.

this dissertation is more general in that registration is performed on sets of images, of arbitrary number, and is not constrained by an initial class labeling.

3.1.1 Mutual Information-Based Image Registration

The application of mutual information to image registration involving two scalar images has been well studied. Mutual information, defined in Appendix A, measures the statistical dependence between two random variables, or the amount of information one random variable contains about the other. In the context of image registration, the random variables represent image intensities. Multi-modal images of the same scene, e.g., a human brain, represent measurements of different properties of objects in that scene, e.g., white matter. These image intensities are not typically statistically independent observations of the underlying physical reality. When multi-modal images are aligned, measurements from one modality reduce the uncertainty in others; at alignment, the mutual information between two images is considered to be maximal.

Rigid image registration by maximizing mutual information was first proposed independently by [15] and [103] and extended to affine image registration by [91]. Mutual information-based image registration can generally be defined as finding the transformation that maximizes the mutual information similarity criterion subject to some geometric constraint on that transformation, typically smoothness. For example, con-

sider the problem of finding a transformation, h, that maps a moving image I_m into the space of a fixed image I_f. Let X and Y be the random variables associated with the image intensities in I_m and I_f respectively. Estimates of the joint and marginal distributions, $p_{fm}(i_f, i_m)$, $p_f(i_f)$, and $p_m(i_m)$, required by mutual information can be obtained by simply normalizing the joint and marginal histograms of the overlapping parts of both I_m and I_f. These techniques are, naturally, sensitive to sampling issues. The intensities i_f and i_m are related by the transformation h. The mutual information similarity criterion states that I_f and I_m are aligned by the transformation $\hat{h}$, for which $I(X;Y)$ is maximal:

$$\hat{h} = \underset{h}{\operatorname{argmin}} \sum_{i_f} \sum_{i_m} p_{f,m}(i_f, i_m) \log \frac{p_{f,m}(i_f, i_m)}{p_f(i_f) p_m(i_m)}.$$

Multi-modal image registration using mutual information is particularly attractive in that no image preprocessing other than computing the marginal and joint distributions is required and there are no functional assumptions regarding the relationship between the intensities in both modalities. This suggests an automated approach where neither segmentation nor landmarks are required. The mutual information similarity criterion does, however, suffer from a few weaknesses: no geometric prior information is incorporated, no assumptions about the data (e.g., known intensity relationships between modalities for certain structures) are incorporated, and there are sampling issues to be addressed (e.g., probability distributions are derived from discrete histograms, the smaller the scale the less robust the histograms).

A comprehensive survey of image registration via mutual information maximization is given in [76]. The authors cover the major considerations: preprocessing, similarity measure (entropy definition choice, normalization, and whether computed locally or globally), transformation group (rigid, affine, and non-linear), implementation (probability distribution estimation and optimization techniques), image dimensionality (e.g., 2D/3D and 3D/3D), number of images, modalities (mono-, multi-, and modality to model), and applications to various human organs. High-dimensional image registration in the context of mutual information and other dissimilarity measures has also been studied. A thorough investigation of these dissimilarity measures in high-dimensional image registration is presented in [42].

A number of methods have been developed to refine the mutual information-based image registration further through generalized versions of Shannon entropy, see [77] for

a review. The most common extension is based on Rényi-Entropy [81, 113]. These divergence measures are typically parameterized by a single continuous variable whose optimization is believed to provide more accurate registration, through control of the measurement sensitivity of the joint histogram, and faster convergence than traditional mutual information [40, 74, 64, 65].

3.2 Multi-Modal Image Set Registration

This subsection examines two topics directly related to the multi-modal image set registration method presented in this dissertation: multiple-image registration methods, beginning with multivariate mutual information, and joint segmentation and registration methods. In the final portion of this subsection, multi-modal image set registration is presented.

3.2.1 Multivariate Mutual Information in Image Registration

As traditional mutual information has successfully been applied to the problem of scalar multi-modal image registration, it is natural to consider applying multivariate mutual information, defined in Section **??**, to the problem of multi-modal image set registration. To that end, consider the redundancy interpretation of multivariate mutual information, defined in Section A.5.2. The scalar multi-modal registration example of Section 3.1.1 is extended as follows: consider the problem of finding a transformation h that maps a moving image set $\bar{I}_m$ into the space of a fixed image set $\bar{I}_f$. Let $\bar{X} = \{X_i\}_{i=1\ldots F}$ and $\bar{Y} = \{Y_i\}_{i=1\ldots M}$ be the random vectors associated with image intensities in $\bar{I}_f$ and $\bar{I}_m$ respectively. Further let $p_f\left(i_{f_1}, \ldots, i_{f_F}\right)$, $p_m\left(i_{m_1}, \ldots, i_{m_M}\right)$, and $p_{fm}\left(i_{f_1}, \ldots, i_{f_F}, i_{m_1}, \ldots, i_{m_M}\right)$ be the probabilities associated with $\bar{X}$ and $\bar{Y}$. The image intensities $\{i_f\}$ and $\{i_m\}$ are related by the transformation h. The multivariate mutual information (redundancy interpretation) criterion states that $\bar{I}_f$ and $\bar{I}_m$ are brought into register by the transformation $\hat{h}$ for which the following is maximized:

$$\begin{aligned}\hat{h} = \underset{h}{\operatorname{argmin}}\, & H\left(p_f\left(i_{f_1}, \ldots, i_{f_F}\right)\right) + H\left(p_m\left(i_{m_1}, \ldots, i_{m_M}\right)\right) \\ & - H\left(p_{fm}\left(i_{f_1}, \ldots, i_{f_F}, i_{m_1}, \ldots, i_{m_M}\right)\right).\end{aligned}$$

This image set registration formulation is problematic in that it requires maintaining enormous sparsely populated joint histograms [9]. Most entries in the joint

histogram will be populated with a zero or one, resulting in a very flat histogram. Thus, multivariate mutual information should give a constant value for many transformations, leading to many local minima in the optimization cost function. Consider a multi-modal image set registration involving four twelve-bit DICOM (Digital Imaging and COmmunications in Medicine) images, an example of which is provided in Section 3.3.2. Using multivariate mutual information would require the construction of a $2^{4 \cdot 12} \approx 2.8 \times 10^{14}$-bin joint histogram, which is impractical. The number of spatial elements for each image is $256 \times 256 \times 170 \approx 1.1 \times 10^7$ voxels yielding an average bin count of $\frac{1.1 \times 10^7 \text{ voxels}}{2.8 \times 10^{14} \text{ bins}} \approx 4 \times 10^{-8}$ voxels/bin. Given these difficulties, this dissertation introduces a model-based approach where the registration is performed using underlying anatomical structures. These anatomical structures are incorporated as a prior in a Bayesian framework. Before describing the proposed framework, several several related methods for multiple image registration are considered.

3.2.2 Existing Multiple Image Registration Methods

To address the challenges posed by extending mutual information to multiple image registration, several groups have developed different approaches. An early example of multivariate mutual information image registration involves co-registration of a time series of PET images illustrating various stages of a radio tracer uptake [2]. This is an example of the use of dissimilar images within the same imaging modality. The method simplifies the joint histogram problem through data reduction via principal component analysis to produce eigen-images. Another approach incorporates local spatial relationships by extending mutual information to consider a neighborhood of pixels around each point in the image space [85]. The idea is to use local spatial information in a traditionally global statistical approach. The authors make the simplifying assumptions that the intensity distributions are multivariate Gaussians and are independent. The independence assumption allows for entropy to be summed along each direction in feature space. A recently introduce method uses linear combinations of conditional entropies to drive the registration [110, 111]. Upper bounds on the the conditional entropies are developed based on joint histograms of lower dimensionality. All of these multivariate methods have been developed in the linear registration setting. The method presented in this dissertation is unique among these in that it is the only algorithm that performs fully 3D non-rigid multi-modal image set registration and does not require computing joint histograms.

3.2.3 Existing Joint Segmentation and Registration Methods

As the proposed multi-modal image set registration framework, described in the next section 3.2.4, involves the joint estimation of segmentation and registration, similar current methods are briefly reviewed. Central to many of these methods is the consideration of local structure which leads to more robust registration and, hence, improved segmentation. A minimax entropy-based registration framework to simultaneously and iteratively segment 2D portal images and register them to 3D CT data is presented in [6]. This method relaxes the independent and identically distributed image pixel intensity assumption associated with mutual information by incorporating correlation among neighboring pixels. Another approach involves a geometric, variational framework that uses active contours to simultaneously segment and register multiple features [109]. Multiple images are segmented by evolving a single contour as well as mapping that contour to each image. A distinctly different approach builds upon the previous method by using logic models while relaxing the requirement that images be completely registered [72]. Markov random fields have been applied in Bayesian maximum aposteriori model segmentation estimation and registration [106]. A Markov random field model-based approach that incorporates a pixel attribute vector based on a pharmacokinetic model is presented in [108, 107]. This method uses a Gaussian data likelihood model and provides spatial coherence and smoothness through the use of a prior. A framework that combines segmentation, bias field correction, and registration into a generative approach is presented in [3]. This method also uses a Gaussian data likelihood model. Iterated conditional modes, introduced by [8], is used to minimize their mixture objective function where each iteration involves alternating between estimating different groups of parameters while holding the others fixed. The method most similar to the one proposed in this dissertation incorporates tissue class information into non-rigid registration by using Kullback-Leibler divergence as a similarity measure between ideal and actual joint voxel class distribution matrices [20]. These matrices represent local fuzzy correspondence of object class labels for pairs of corresponding voxels. The method presented in this dissertation differs from most of these approaches in that it involves fully 3D non-rigid registration.

3.2.4 Multi-Modal Image Set Registration

Given the aforementioned difficulties with using multivariate mutual information in the image set registration setting, a novel model-based approach is proposed where

the registration is performed using underlying anatomical structures. These structures are incorporated as a geometric prior in a Bayesian framework. From the theory of pattern classification [21], it is known that the use of a Bayesian classifier for decision making is optimal in that it corresponds to minimizing the average probability of error associated with the decision. Thus, Bayesian classification using a geometric prior provides a goodJos approach for generating class-conditional densities that describe the anatomical structures.

This framework is based on the assumption that human brain anatomy consists of finitely enumerable structures such as grey matter, white matter, and cerebrospinal fluid. These structures present with varying radiometric intensity values across disparate imaging modalities. Given two multi-modal image sets, the underlying structures are captured by estimating, for each image set, the class-conditional posterior maps associated with each structure. These class posteriors are then used to produce a coordinate-independent average posterior by estimating dense diffeomorphic registration maps relating the domains of the two class posteriors. The sum-of-Kullback-Leibler divergences $\bar{D}\left(\hat{p}||\{p_i\}_{i=1}^{N}\right)$ with $N = 2$ distributions is used as a distance function on the space of probability mass functions to estimate the transformations. The use of class posteriors provides an image intensity-independent approach to image registration.

Specifically, consider the problem of finding a mapping between image sets $\bar{I}_1$ and $\bar{I}_2$ (Figure 3.3). That is, find the mappings $f : \Omega_1 \rightarrow \Omega_2$ and $g : \Omega_2 \rightarrow \Omega_1$ where Ω_1 and Ω_2 are the domains of image sets $\bar{I}_1$ and $\bar{I}_2$ respectively. To facilitate the registration, a new domain Ω, independent of Ω_1 and Ω_2 is introduced. Let transformations h_1 and h_2 map Ω to Ω_1 and Ω_2 respectively. By construction, $f = h_2 \circ h_1^{-1}$ and $g = h_1 \circ h_2^{-1}$. This registration method is inverse consistent as $f \circ g = g \circ f = e$, the identity map. Having described the transformation framework the Bayesian framework for representing the anatomical class structures is presented next.

3.2.4.1 Bayesian Framework

The underlying neuroanatomy, represented in two acquired sets of multi-modal images, is assumed to consist of a set, C, of separate anatomical structure classes, c_j. From the multi-modal image sets $\bar{I}_1$ and $\bar{I}_2$, for each class $c_j \in C$ jointly estimate the posterior mass functions $p_1(x) = p(c_j(h_1(x))|\bar{I}_1)$ and $p_2(x) = p(c_j(h_2(x))|\bar{I}_2)$ along with the registration maps $h_1(x)$ and $h_2(x)$, that map the independent domain $\Omega \subset \mathbb{R}^3$, into the domain of $\bar{I}_1$, $\Omega_1 \subset \mathbb{R}^3$, and $\bar{I}_2$, $\Omega_2 \subset \mathbb{R}^3$, respectively. This method is independent of the number of images comprising each image set. Optimal inter-subject

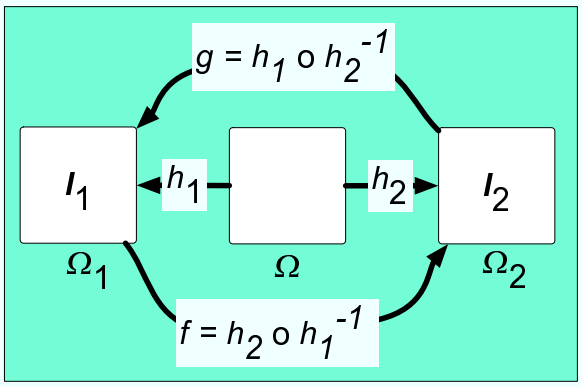


Figure 3.3: Inverse-Invariant Registration

Registration of image sets $\bar{I}_1$ and $\bar{I}_2$ through the unbiased domain Ω.

multi-modal image set registration is estimated by an alternating iterative algorithm which is motivated by an expectation maximization method used in [80, 102]. The proposed algorithm interleaves the estimation of the posteriors associated with $\bar{I}_1$ and $\bar{I}_2$ and the estimation of the registration maps $h_1 : \Omega \rightarrow \Omega_1$ and $h_2 : \Omega \rightarrow \Omega_2$. Figure 3.4 depicts this Bayesian framework.

For each class c_j, the associated data likelihoods $p(\bar{I}_{\{1,2\}}(x)|c_j(x), \mu_j, \Sigma_j)$, are modeled as multivariate normal distributions with means μ_j, and covariances Σ_j. Given the transformations h_1 and h_2 and the current estimates μ_j and Σ_j for both image sets, the posteriors of $\bar{I}_1$ and $\bar{I}_2$ are associated with the independent coordinate probability mass function p_Ω by using Bayes's Rule with p_Ω as the prior for both posteriors $p_1(x)$ and $p_2(x)$. Having defined the posteriors, the parameters μ_j and Σ_j are updated by their expected values. An alternative approach is the non-parametric kernel density estimation described in [79].

3.2.4.2 Large Deformation Diffeomorphic Registration

With the model-based Bayesian framework defined, now consider the problem of estimating a class posterior average, $\hat{p}$, from p_1 and p_2, representing image sets $\bar{I}_1$ and $\bar{I}_2$ respectively. The average $\hat{p}$ is neither p_1 nor p_2. Consider the problem of constructing a mapping between $\hat{p}$ and each of p_1 and p_2. That is, estimate the mappings $h_1 : \Omega \rightarrow \Omega_1$ and $h_1 : \Omega \rightarrow \Omega_2$ where $\Omega \subset \mathbb{R}^3$, $\Omega_1 \subset \mathbb{R}^3$and $\Omega_2 \subset \mathbb{R}^3$ are the domains of the class posteriors $\hat{p}$, p_1 and p_2 respectively. The domain Ω is chosen to be independent of the

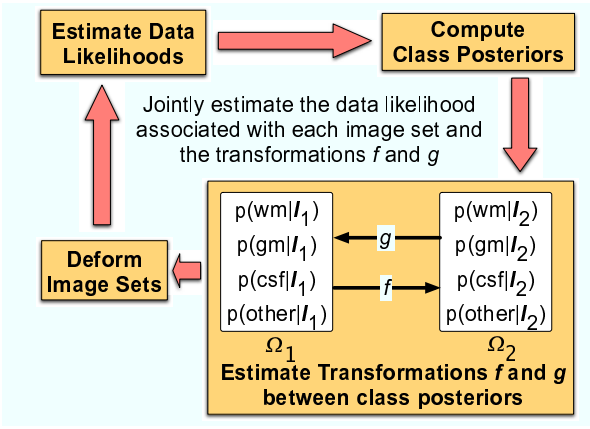


Figure 3.4: Model-Based Image Set Registration

Registration of image sets $\bar{I}_1$ and $\bar{I}_2$ through the unbiased domain Ω.

individual subject class posterior domains, Ω_1 and Ω_2. This framework is depicted in Figure 3.5.

The desired average class posterior $\hat{p}$ is the one that requires the minimum amount of energy to be deformed into both p_1 and p_2. More precisely, given a transformation group S with associated metric $D : S^2 \to \mathbb{R}$, along with a probability density dissimilarity measure $E(p, q)$, to find the class posterior map $\hat{p}$ such that

$$\{\hat{h}_{\{1,2\}}, \hat{p}\} = \underset{h_{\{1,2\}} \in S, p}{\operatorname{argmin}} \left[E(p_1 \circ h_1, p) + E(p_2 \circ h_2, p) + D(e, h_1)^2 + D(e, h_2)^2 \right] \tag{3.1}$$

where $e(x) = x$ is the identity transformation.

Within the framework of computational anatomy a commonly used transformation group is the group of diffeomorphisms [48]. Central to this framework, therefore, is the assumption of homogeneous anatomy between subjects [32]. As diffeomorphic maps preserve topology they are ideal for the quantitative study of shape since they neither bend nor tear space. The fluid image registration method used in this dissertation is motivated by the image matching problem formulated via fluid flows introduced by [14] and posed a control formulation in [22]. In this formulation, the optimal diffeomorphic match is constructed to minimize a running smoothness cost on the velocity field generating the diffeomorphism, further described below, while simultaneously minimizing

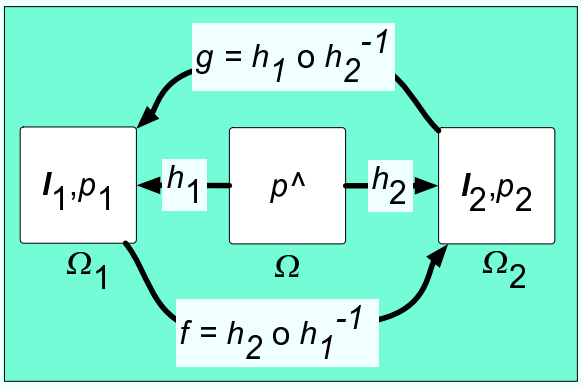


Figure 3.5: Image Registration Through Class Posteriors

Associated with each image set $\bar{I}_i$ is a class-conditional probability map p_i and is related to the unbiased domain Ω through the transformation h_i.

an image dissimilarity term. A gradient descent approach is taken where a variational cost is measured in terms of a velocity field.

In this dissertation, the class-posterior map matching involves the particular transformation h that maps the space of a class-posterior representing a subject, p_i, to the space of the class-posterior map representing the average, $\hat{p}$, $p_i(h_i(x)) = \hat{p}(x)$. In the large deformation setting, a given h is estimated as the end point point of the flow associated with a smooth time-dependent vector field. More specifically, the approach is to construct diffeomorphisms $h : \Omega \subset \mathbb{R}^3 \leftrightarrow \Omega$ in terms of solutions to the ordinary differential equation defined by the nonlinear transport equation

$$\frac{d}{dt}h(x,t) = v(h(x,t),t) \quad t \in [0,1] \tag{3.2}$$

with boundary condition,

$$h(x,0) = x \quad x \in \Omega$$

as presented in [70, 48]. The boundary condition corresponds to no deformation. The solution to Equation 3.2 is the function $h(x,t)$ that satisfies

$$h(x,t) = h(x,0) + \int_0^t v(h(x,\tau),\tau)d\tau. \tag{3.3}$$

The final time diffeomorphism $h(x,1)$ mapping the anatomy is therefore controlled via the velocity field $v(\cdot,t)$, $x \in \Omega$, $t \in [0,1]$,

$$h(x,1) = x + \int_0^1 v(h(x,\tau),\tau)d\tau.$$

Before discussing how the diffeomorphic transformations are generated, it is important to know that they do exist in this setting. In his dissertation, Joshi proves the following theorem.

Theorem 3.1 (Diffeomorphism Existence)**.** *Let $\Omega = [0,1]^3 \subset \mathbb{R}^3$ and $v : (x,t) \in \Omega \times [0,1] \rightarrow v(x,t) \in \mathbb{R}^3$ be a continuously differentiable vector field with compact support contained in Ω for each $t \in [0.1]$. Let h be the solution to the system of ordinary differential equations*

$$\frac{d}{dt}h(x,t) = v(h(x,t),t)$$

with the initial condition $h(0,x) = x$. Then for each $t \in [0,1]$, $h(\cdot,t)$ is a diffeomorphism of $\Omega \leftrightarrow \Omega$.

Proof. See [48]. □

The boundary condition $h(x,0) = x$ represents the Lagrangian interpretation of fluid flow. The existence of diffeomorphisms under the Eulerian interpretation of fluid flow where $h(x,1) = x$ is shown in [22]. Conditions are formulated under which the regularity of $v(\cdot,t)$ imposed by finiteness of the the norm $||v||^2$ guarantees that the associated flow, h, is supported on the space of diffeomorphisms. The key issue is the choice of norm $||\cdot||^2$.

Following [7], one choice to ensure existence of solutions in the space of diffeomorphisms in Equation 3.2 has been to construct V as the completion of the space of smooth, compactly-supported vector fields for the inner-product defined through a differential operator L given by

$$\langle f,g\rangle_V \doteq \langle Lf,Lg\rangle_2 = \langle L^\dagger Lf,g\rangle_2 \quad f,g \in V \tag{3.4}$$

where $L^\dagger$ denotes the adjoint of L, and $\langle\cdot,\cdot\rangle_2$ is the usual L^2-product for square inte-

grable vector-fields on Ω. That is,

$$\int_0^1 ||v(\cdot,t)||_V^2 dt = \int_0^1 ||L^\dagger L v(\cdot,t)||_2^2 dt.$$

With V defined in this way, the flow $v \in L^1(V,[0,1])$ generates the sub-group of diffeomorphisms $G \doteq \{\phi \in V | \phi(h(x,1),1), \quad v \in L^1(V,[0,1])\}$ [22]. From these assumptions on V, a compact self-adjoint operator $K : L^2(\Omega,\mathbb{R}^d) \rightarrow V$ is uniquely defined by $\langle x,y\rangle_2 = \langle Kx,y\rangle_V$. That is, any smooth vector field $f \in V$ can be obtained such that

$$K(L^\dagger L)f = f.$$

In summary, the norm $||v(\cdot,t)||_V$ generates an inner product on the space of continuous vector fields $C^\infty(\Omega)^3$, and $||v(\cdot,t)||_V$ needs to be finite to ensure regularity in each component of v. Smoothness in v is induced by adding $||Lv(\cdot,t)||^2$ as a regularizing cost penalty term in the optimization.

The choice of operator L is typically governed by the notion of a prior, in the Bayesian sense, on the transformations obtained by driving a fluid or elastic media or by a noise process. In such a case L is determined by the statistics of the noise and the constitutive laws of the media. Following [14], in this dissertation a modified Navier-Stokes operator $L = \alpha\nabla^2 + \beta\nabla(\nabla\cdot) + \gamma$ is used where pressure gradient and inertial terms are neglected and a low Reynold's number is assumed.

A differential operator L is completely characterized by its Green's function $G(x,y)$ where $LG(x,y) = \delta(x-y)$. Under the noise process model, the induction of prior through constant coefficient local (bounded support), positive definite (invertible) differential operator L satisfying $Lu(x) = e(x)$ where $e(x)$ is white noise. Then, as shown in [52], $u(x)$ is a zero-mean Gaussian process with covariance

$$K(x,y) = \int G(x,u)G(y,u)du.$$

Throughout assume the compact setting $\Omega = [0,1]^3$ with the operator and boundary conditions chosen so that the Green's function is non-singular and continuous in both variables so that the matrix $K(x,y) = G(x,y)G(x,y)^\dagger$ is positive definite as an operator.

Having chosen a linear differential operator to define the inner-product $\langle\cdot,\cdot\rangle_V$ and having established the conditions under which the solution to Equation 3.2 generates diffeomorphic transformation, the image registration optimization problem in Equation

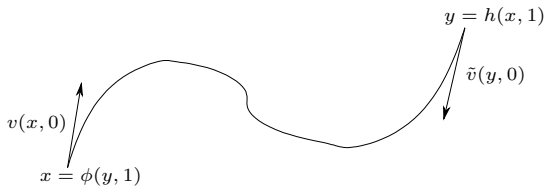


Figure 3.6: Velocity Fields

3.1 is further specified. The transformations h_i are generated by integrating velocity fields forward in time, and h_i^{-1} are generated by integrating velocity fields backward in time. The relationship between spatial locality, velocity fields, and time is shown in Figure 3.6. The spatial location y is described in terms of the forward integration of the velocity field v starting from spatial location x. That is,

$$y = h(x, 1) = x + \int_0^1 v(h(x, \tau), \tau) d\tau.$$

Similarly, x can be described in terms of integrating the reverse velocity field $\tilde{v}$ starting at y. That is,

$$x = \phi(y, 1) = y + \int_0^1 \tilde{v}(\phi(y, \tau), \tau) d\tau.$$

From Figure 3.6, note that

$$v(h(x, t), t) = -\tilde{v}(\phi(y, 1 - t), 1 - t)$$

and hence

$$||Lv(x, t)||^2 = ||L\tilde{v}(y, 1 - t)||^2$$

where

$$L = \alpha\nabla^2 + \beta\nabla(\nabla\cdot) + \gamma$$

is, again, a modified Navier-Stokes operator.

The metric on the space of diffeomorphisms is induced using a Sobolev norm via the partial differential operator L on the velocity fields v. Let h be a diffeomorphism isotopic to the identity transformation e. Define the squared distance $D^2(e, h)$ as

$$D^2(e,h) = \min_v \int_0^1 \int_\Omega ||Lv(x,t)||^2 dxdt \tag{3.5}$$

subject to

$$h(x) = x + \int_0^1 v(h(x,t),t)dt.$$

The distance between any two diffeomorphisms is defined by

$$D(h_1,h_2) = D(e, h_1^{-1} \circ h_2).$$

The construction of h and h^{-1}, as well as the properties of D, are described in [68, 71].

This distance satisfies all of the properties of a metric. Namely it is non-negative, symmetric, and satisfies the triangle inequality. The distance D is trivially non-negative. Symmetry follows from the fact that h^{-1} is generated by integrating backwards in time the negative of the velocity field that generates h. Hence the minimizer is the same for both h and h^{-1}, implying that $D(e,h) = D(e,h^{-1})$. A detailed discussion of D, including a demonstration of how it satisfies the triangle inequality, is given in [71].

Having defined a metric on the space of diffeomorphism and a regularization operator L, the energy minimization problem described in Equation 3.1 is formulated as

$$\begin{aligned}\{\hat{h}_{\{1,2\}}, \hat{p}\} = \underset{h_{\{1,2\}},p}{\operatorname{argmin}} & E(p_1 \circ h_1, p) + E(p_2 \circ h_2, p) \\ & + \int_0^1 \int_\Omega ||Lv_1(x,t)||^2 dxdt + \int_0^1 \int_\Omega ||Lv_2(x,t)||^2 dxdt \end{aligned} \tag{3.6}$$

subject to

$$h_i(x) = x + \int_0^1 v_i(h_i(x,t),t)dt.$$

3.2.4.3 Registration

At a given spatial location $x \in \Omega$, the dissimilarity between image sets $\bar{I}_1(x)$ and $\bar{I}_2(x)$ is measured by the dissimilarity between the posterior mass functions modeling them, $p_1(x)$ and $p_2(x)$. Minimizing the $\bar{D}\left(\hat{p}||\{p_i\}_{i=1}^N\right)$ between p_1 and p_2 maximizes a lower bound on the Bayes' probability of error and thus renders the probability mass

functions more indistinguishable. That is, it brings them closer together. The following distance, using Equation 2.10 with $N = 2$, is used to drive the registration at position $x \in \Omega$:

$$\begin{aligned} E(p_1(x), p_2(x)) &= \bar{D}\left(\hat{p}(x)||\{p_1(x), p_2(x)\}\right) \\ &= -2\log \sum_{c_j \in C} \left(p_1(h_1(c_j(x))|\bar{I}_1) p_2(h_2(c_j(x))|\bar{I}_2)\right)^{\frac{1}{2}}. \end{aligned}$$

With this result, the minimization problem stated in Equation 3.1 is rewritten as follows:

$$\begin{aligned} \hat{v}_1, \hat{v}_2 = \underset{v_1, v_2}{\operatorname{argmin}} & \int_\Omega \log \sum_{c_j \in C} \left(p_1(h_1(c_j(x))|\bar{I}_1) p_2(h_2(c_j(x))|\bar{I}_2)\right)^{\frac{1}{2}} dx \\ & + \int_0^1 \int_\Omega ||Lv_1(x)||^2 dx dt + \int_0^1 \int_\Omega ||Lv_2(x)||^2 dx dt. \end{aligned}$$

3.2.4.4 Implementation

The method proposed in this dissertation uses Christensen's greedy algorithm for propagating templates [14]. The variation for h_1 of the average $E(p_1(x), p_2(x))$ term is computed as

$$\begin{aligned} \frac{\partial}{\partial h_1} \frac{1}{|\Omega|} \int_\Omega \bar{D}\left(\hat{p}(x)||\{p_1(x), p_2(x)\}\right) dx &= -\frac{2}{|\Omega|} \frac{\partial}{\partial h_i} \log \sum_{c_j \in C} \left(p_1(c_j(x)) p_2(c_j(x))\right)^{\frac{1}{2}} dx \\ &= -\frac{2}{|\Omega|} \cdot \int_\Omega \frac{\sum_{c_j \in C} \frac{\partial}{\partial h_1} \left(p_1(c_j(x)) p_2(c_j(x))\right)^{\frac{1}{2}}}{\sum_{c_k \in C} \left(p_1(c_k(x)) p_2(c_k(x))\right)^{\frac{1}{2}}} dx \\ &= -\frac{1}{|\Omega|} \int_\Omega \frac{\sum_{c_j \in C} \left(\frac{p_2(c_j(x))}{p_1(c_j(x))}\right)^{\frac{1}{2}} \nabla p_1|^T_{c_j(h_1(x))}}{\sum_{c_k \in C} \left(p_1(c_k(x)) p_2(c_k(x))\right)^{\frac{1}{2}}} dx. \end{aligned}$$

The variation for h_2 is computed in a similar manner. The velocity fields $v_{\{1,2\}}$ at each iteration are updated by solving the partial differential equations

$$Lv_{\{1,2\}}(x,t) = \frac{\partial}{\partial h_{\{1,2\}}} \frac{1}{|\Omega|} \int_{\Omega} \bar{D}\left(\hat{p}(x)||\{p_1(x), p_2(x)\}\right) dx.$$

The velocity fields v_i are computed at each iteration by applying the inverse of the differential operator L to the body force function $f_i(x) = \frac{\partial}{\partial h_i} \frac{1}{|\Omega|} \int_{\Omega} \bar{D}\left(\hat{p}(x)||\{p_1(x), p_2(x)\}\right) dx$, i.e. $v_i(x) = L^{-1} f_i(x)$. This computation is performed in the Fourier domain [50] using the Fast Fourier transform.

The forward and inverse integration are described as follows. At time t, the transformations h_i are described as

$$\begin{aligned} h_i(x, t+\delta) &= h_i(x,t) + \int_t^{t+\delta} v_i(h_i(x,\tau),\tau) d\tau \\ &\approx h_i(x,t) + \delta v_i(h_i(x,t),t) \end{aligned}$$

for small δ. At iteration k of the algorithm, the transformations h_i become the telescoping compositions $h_i = h_i^k \circ h_i^{k-1} \circ \cdots \circ h_i^1$. At time t, the inverse transformations h_i^{-1} are described as

$$\begin{aligned} h_i^{-1}(y,t) &= h_i^{-1}(y - \int_t^{t-\delta} v_i(y,\tau) d\tau, t-\delta) \\ &\approx h_i^{-1}(y - \delta v_i(y,t), t-\delta) \end{aligned}$$

for small δ. At iteration k of the algorithm, the transformations h_i^{-1} become the telescoping compositions $h_i^{-1} = h_i^{-1,1} \circ h_i^{-1,2} \circ \cdots \circ h_i^{-1,k}$.

Appendix B discusses the likelihood of the actual optimum in the multi-modal image set registration cost function being achieved.

3.3 Results

To evaluate the image set registration method, a collection of image sets of increasing complexity was defined. For each image subject, the image sets were created using 3D scalar images from a population of four imaging modalities: MRA, T1-FLASH MR, T1-MPRAGE MR, and T2 MR. The composition of these image sets is described in Section 3.3.2.1. The individual MR images were acquired at UNC Chapel Hill using

a Siemens head-only 3-Tesla system (Allegra, Siemens Medical Systems Inc.) and a Siemens 1.5-Tesla system (Sonata, Siemens Medical Systems Inc.) with a head coil. Imaging parameters for the T1 and T2 image acquisitions are as follows, for the T1 images, a TR of 15msec, a TE of 7msec, a TH of 1mm, and an in-plane resolution of 1×1 mm^2 and for the T2 images, a TR of 7730msec, a TE of 80msec, a TH of 1mm, and an in-plane resolution of 1×1 mm^2. Additionally, a 3D time-of-flight MRA sequence was acquired. Velocity compensation along both the frequency and the phase encoding direction was used to maximize signal de-phasing induced by the flowing spins. A magnetization transfer pulse was used to suppress signal from brain parenchyma while maintaining signal from flowing spins. The acquired voxel spacing for the MRA images was $0.5134 \times 0.5134 \times 0.78$ mm^3 and $1 \times 1 \times 1\,$mm^3 for the T1 and T2 images. The MRA images were resampled to $1 \times 1 \times 1\,$mm^3.

3.3.1 Data Preprocessing

The tissue exterior to the brain was removed using a mask generated by a brain segmentation tool based on the statistical classification method described in [79]. The geometric prior used to initialize the algorithm was also produced using this tool. Mid-axial, mid-coronal, and mid-sagittal slice views for subjects 1 and 2 are presented in Figures 3.7 and 3.8 respectively. These four modalities provide complementary information. For example, the T1-FLASH and T1-MPRAGE images have contrast differences, and the MRA images exhibit missing information due to grey matter/white matter washout and axial slab effect. In these examples, the set of structural classes is taken to be

$C = \{c_1 = \text{grey matter}, c_2 = \text{white matter}, c_3 = \text{cerebrospinal fluid}, c_4 = \text{other}\}$.

3.3.2 Registration Experiments

To evaluate this image set registration framework, the transformations, f_1 and g_1, relating the domains of subject 1 and subject 2 were estimated by applying the multi-modal image set method to a mono-modal registration. These two transformations were then used as "ground truth" for the purpose of evaluating an increasingly complex collection of image set registrations. Inverse-consistency error was computed to quantitatively evaluate the results.

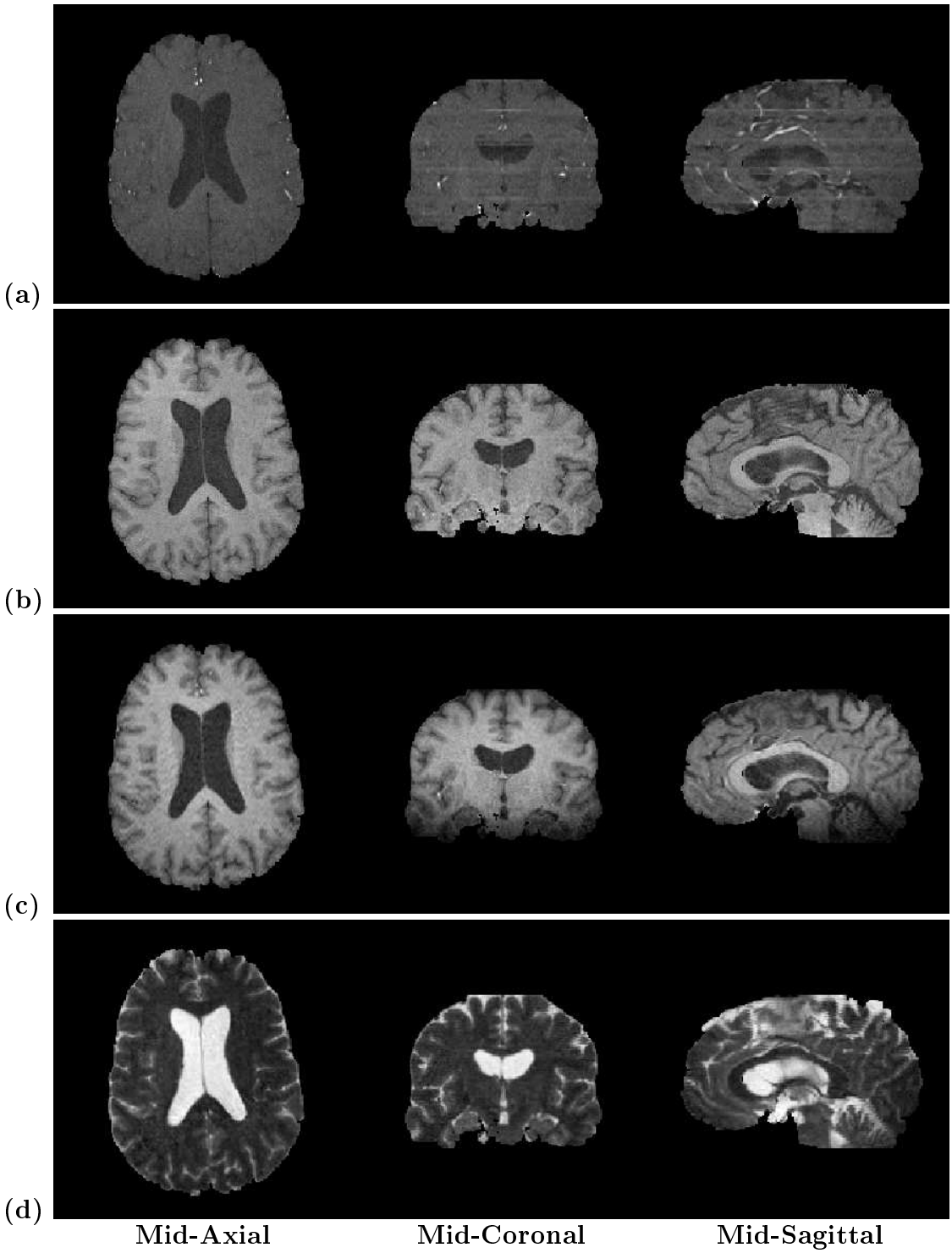


Figure 3.7: Subject One

The orthogonal slice views into the four multi-modal scalar images for subject one: MRA (a), T1-FLASH (b), T1-MPRAGE (c), and T2 (d).

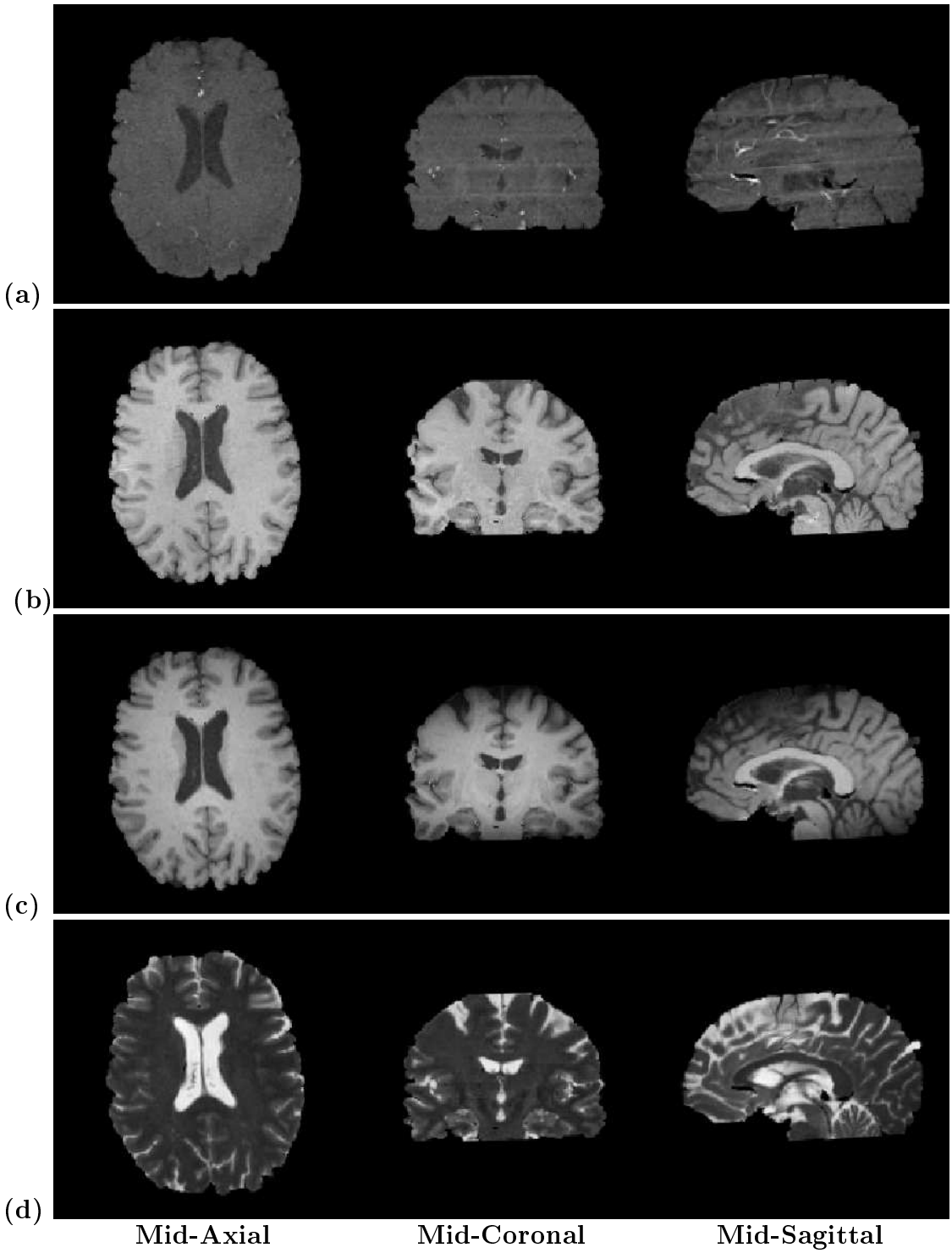


Figure 3.8: Subject Two

The orthogonal slice views into the four multi-modal scalar images for subject two: MRA (a), T1-FLASH (b), T1-MPRAGE (c), and T2 (d).

3.3.2.1 Setup

The following eight registration experiments were performed:

1. Mono-modal/Mono-modal (common): $\bar{I}_1$ = T1-FLASH of subject 1 and $\bar{I}_2$ =T1-FLASH of subject 2.

2. Mono-modal/Mono-modal (mutually exclusive): $\bar{I}_1$ = T1-FLASH of subject 1 and $\bar{I}_2$ =T2 of subject 2.

3. Bi-modal/Bi-modal (fully common): $\bar{I}_1$ = T1-FLASH and T2 of subject 1 and $\bar{I}_2$ = T1-FLASH and T2 of subject 2.

4. Bi-modal/Bi-modal (single common): $\bar{I}_1$ = T1-FLASH and T2 of subject 1 and $\bar{I}_2$ = T1-MPRAGE and T2 of subject 2.

5. Bi-modal/Bi-modal (mutually exclusive): $\bar{I}_1$ = T1-FLASH and T2 of subject 1 and $\bar{I}_2$ = T1-MPRAGE and MRA of subject 2.

6. Bi-modal/Mono-modal (mutually exclusive): $\bar{I}_1$ = T1-FLASH and T2 of subject 1 and $\bar{I}_2$ = MRA of subject 2.

7. Tri-modal/Tri-modal (fully common): $\bar{I}_1$ = T1-FLASH, T1-MPRAGE, and T2 of subject 1 and $\bar{I}_2$ = T1-FLASH, T1-MPRAGE, and T2 of subject 2.

8. Quad-modal/Quad-modal (fully common): $\bar{I}_1$ = T1-FLASH, T1-MPRAGE, T2, and MRA of subject 1 and $\bar{I}_2$ = T1-FLASH, T1-MPRAGE, T2, and MRA of subject 2.

From each of these experiments, transformations f_i and g_i were obtained. The first experiment provides the "ground truth" transformations, f_1 and g_2. The T1-FLASH modality was chosen for the first experiment due to its relatively good white matter/grey matter contrast.

3.3.2.2 Bi-Modal/Bi-Modal (Mutually Exclusive) Registration

For the purposes of brevity, only qualitative results for the most interesting of these experiments, the bi-modal/bi-modal mutually exclusive registration are presented. In this experiment, $\bar{I}_1$ represents the T1-FLASH and T2 images acquired from subject one and $\bar{I}_2$ represents the T1-MPRAGE and MRA images acquired from subject two. The

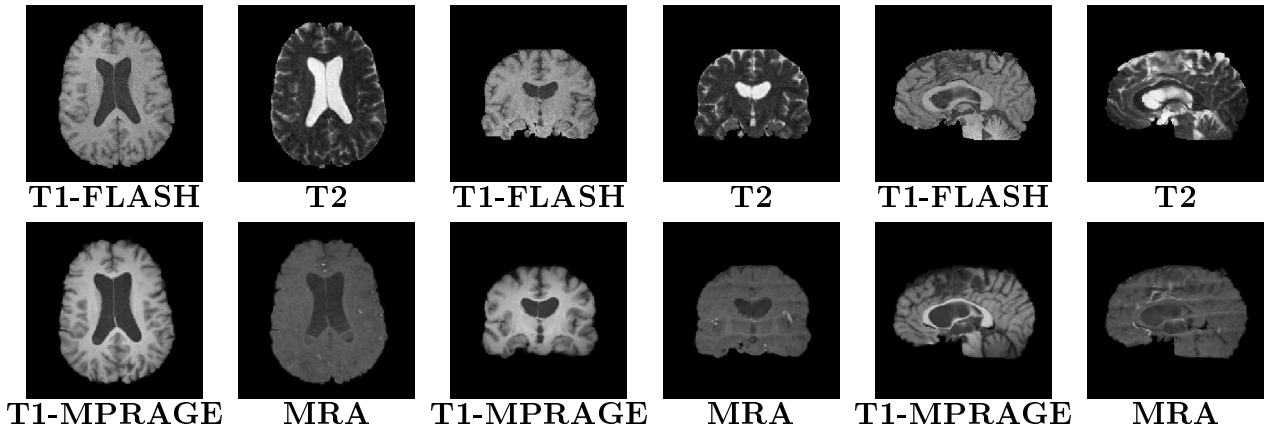


Figure 3.9: Forward Mapping

The top row shows mid-axial, mid-coronal, and mid-sagittal views of image set $\bar{I}_1$. The bottom row shows the same views for the deformed image set $\bar{I}_2 \circ f$.

estimated forward, f, and inverse, g, transformations are depicted, in three orthogonal views, in Figures 3.9 and 3.10 respectively. In Figure 3.11, a qualitative assessment of the registration is made by examining axial slices of $\bar{I}_1$ and $\bar{I}_2$ in greater detail through a checkerboard pattern. The mismatch between the image sets is clearly evident in the top row of the figure. The second and third rows illustrate the effectiveness of the registration under the estimated forward and inverse transformations respectively.

3.3.2.3 Inverse-Consistent Image Registration

A by-product of using the unbiased domain Ω in the multi-modal image set registration framework is inverse-consistent (or inverse-invariant) image registration. A registration framework is inverse consistent if image ordering does not affect the registration result. Many image registration algorithms are not inverse consistent because their image dissimilarity metrics are computed in the coordinate system of one of the images being registered. The choice of such a reference image can bias the result of the registration. Inverse consistent registration is desired when there is no a priori reason to choose one image over another as a reference image.

In traditional techniques for image registration, solutions may be systematically biased with respect to expanding and contracting regions in the estimated transformation [12]. Existing methods for generating inverse consistent registration approximate

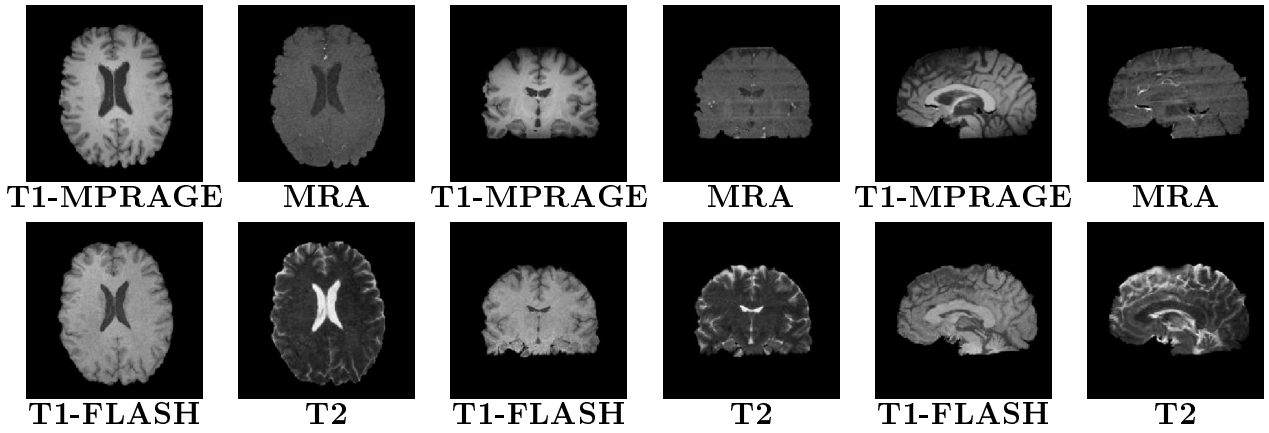


Figure 3.10: Inverse Mapping

The top row shows mid-axial, mid-coronal, and mid-sagittal views of image set $\bar{I}_2$. The bottom row shows the same views for the deformed image set $\bar{I}_1 \circ g$.

inverse consistency by adding an inverse consistency penalty to the optimization cost function. The registration frameworks formulated in these methods are not intrinsically symmetric. Methods for approaching this problem, involving algorithms that estimate incremental transformations while approximating inverse consistency constraints on each incremental transformation, are presented in [34, 66, 39]. The approach presented in this dissertation is intrinsically inverse consistent as the registration problem is formulated symmetrically. Therefore, no correction penalty for consistency is required.

More specifically, a registration, characterized by a transformation, h, is inverse-consistent (symmetric) if, for images I_1 and I_2, the following holds: $h(I_1, I_2) = h^{-1}(I_2, I_1)$. A registration for which this property does not hold is considered to be asymmetric. The following is a list of potential sources for asymmetry in non-rigid registration:

- *Order non-preservation*: Many image registration algorithms are not inverse-consistent since their (dis)similarity metrics are computed in the coordinate system of either one of the images involved in the registration. This leads to order non-preservation of energy cost functions. That is, for an energy function $E(I_1, I_2, h)$ and two estimates, h_1 and h_2, for the transformation h, the following may hold, $E(I_1, I_2, h_1) < E(I_1, I_2, h_2)$ but $E(I_2, I_1, h_1^{-1}) \geq E(I_2, I_1, h_2^{-1})$.

- *Non-linearity in models*: Continuum mechanical methods are used to model de-

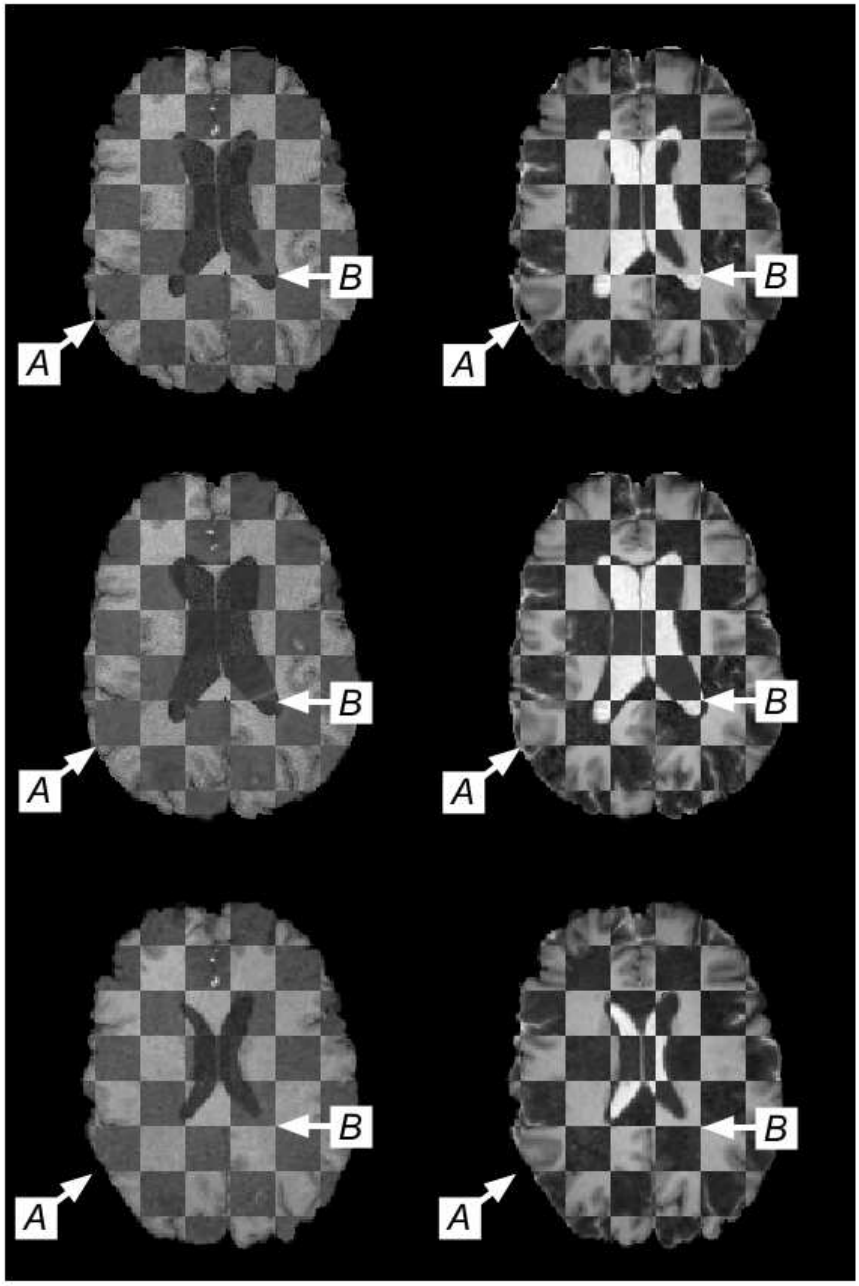


Figure 3.11: Qualitative Assessment

The top row shows a checkerboard blending of an axial slice of the image sets $\bar{I}_1$ and $\bar{I}_2$. The middle row shows a blending between $\bar{I}_1$ and $\bar{I}_2 \circ f$, the forward mapping, and the bottom row shows a blending between $\bar{I}_2$ and $\bar{I}_1 \circ g$, the inverse mapping. The left column is composed of the T1-FLASH of $\bar{I}_1$ and the MRA of $\bar{I}_2$, and the right column is composed of the T2 of $\bar{I}_1$ and the T1-MPRAGE of $\bar{I}_2$. The registration has accommodated the local variability between the two image sets, especially in the cortical region (e.g., point A), and in the ventricular region (e.g., point B).

formations (e.g., thin-plate splines may be non-linear).

- *Non-stable transformation space*: The space of allowable transformations may not form a group, G, and, hence, may not be stable by inversions. That is, $h \in G$ but $h^{-1} \notin G$.

- *Local minima*: The optimization algorithm used for estimating h may get stuck in different local minima of the energy function E when the images $I_1$and I_2 are exchanged.

The L_2 difference norm, $||f_1(g_i(x)) - x||_2$, was used to evaluate inverse consistency between each experiment i, and the first experiment for each spatial location $x \in \Omega$. For numerical stability, these inverse-consistency errors were computed via telescoping compositions as described in Section 3.2.4.4. Over all eight experiments, the maximum computed inverse-consistency error was 3.12×10^{-4} voxels with an average of 5.04×10^{-5} voxels.

Chapter 4

Multi-Class Posterior Atlas Formation[1]

This chapter extends the framework for multi-modal image set registration from Chapter 3 to multi-class posterior atlas formation. Consider the important problem in computational anatomy introduced in Section 1.2, the construction of an exemplar atlas from a population of medical images. Such atlases represent the anatomical variation present in populations [69, 32, 94]. Many images are mapped into a common coordinate system to study intra-population variability and inter-population differences, to provide voxel-wise mapping of functional sites, and to facilitate tissue and object segmentation via registration of anatomical labels. Common techniques for creating atlases often include choosing a template image, which inherently produces a bias. Motivated by the atlas construction framework presented in [49], unbiased multi-class atlases are constructed from populations of anatomical class posteriors using large deformation diffeomorphic registration. When applied to two image sets, this atlas formation method yields the inverse-consistent image set registration of Chapter 3.

Digital atlases of humans are prominent in image segmentation algorithms. For example, a method for automatically segmenting images of normal healthy human brains, based on statistical classification theory [21], is presented in [102, 101]. An extension to human brains with pathology is described in [80]. These methods rely on multi-modal images that have been affinely registered to the coordinate space of the chosen atlas. The registration pre-processing step uses mutual information, described

[1]The work presented in this chapter was done in collaboration with Dr. Sarang Joshi and Brad Davis at the University of North Carolina at Chapel Hill. This chapter is heavily based on previous papers [63, 49].

in [16], as a similarity measure for computing distances between probability mass functions representing joint and marginal product distributions of image intensities. Global transformations, such as affine transformations, are insufficient to accommodate the local variability that exists in any population of human brains. For example, although all humans possess a cerebellum and cortical grey matter enclosing white matter, the location and manner of cortical and cerebellar folding are highly variable across a population.

Most digital atlases involve features that are single numbers (e.g., Hounsfield units in computed tomography (CT) images). Unbiased atlas formation from such scalar images, using the squared-error image dissimilarity measure, is described in [49]. This chapter focuses on multi-class posterior atlas formation from multi-modal image sets. Specifically, the Bayesian framework from Section 3.2.4.1 is applied to a collection of image sets. From a population of N multi-modal image sets $\{\bar{I}_i\}_{i=1}^N$, for each class $c_j \in C$, estimate the class posterior mass functions $p_i(c_j(x)|\bar{I}_i)$ for each image set i where $c_j(x)$ is the class associated with the voxel at spatial position $x \in \Omega \subset \mathbb{R}^3$. This method is independent of the number of images comprising each image set. These class posteriors are produced using the expectation maximization method described in [80, 102]. Following [102], for each class c_j the associated data likelihood, $p(\bar{I}_i(x)|c_j(x), \mu_j, \Sigma_j)$, is modeled as a normal distribution with mean, μ_j, and covariance, Σ_j.

This chapter reviews existing atlas construction methods in Section 4.1. Using the large deformation diffeomorphic framework requires a more general notion of averaging transformations. This generalized notion of averaging to metric spaces is discussed in Section 4.2. The multi-class atlas construction method is then presented in Section 4.3 with a some results comparing affine and diffeomorphic averaging in 4.4.

4.1 Atlas Formation: a Review

Since Brodmann, nearly a century ago, began mapping areas of the cerebral cortex based on cytoarchitectural boundaries [11], the construction of brain atlases has been central to understanding the variability of brain anatomy. More recently, since the advent of modern computing and digital imaging techniques, intense research has been directed towards the development of digital three-dimensional atlases of the brain. Most digital brain atlases so far are based on a single subject's anatomy [43, 104]. Although these atlases provide a standard coordinate system, they are limited because a single anatomy cannot represent faithfully the complex structural variability between

individuals. A major focus of computational anatomy has been the development of image mapping algorithms [29, 68, 84, 98] that can map and transform a single brain atlas onto a population. In this paradigm, the atlas serves as a deformable template [31]. The deformable template can project detailed atlas data such as structural, biochemical, functional, and vascular information onto the individual or an entire population of brain images. The transformations encode the variability of the population under study. A statistical analysis of the transformations can also be used to characterize different populations [18, 44, 92]. For a detailed review of deformable atlas mapping and the general framework for computational anatomy, see [98, 32]. One of the fundamental limitations of using a single anatomy as a template is the introduction of bias, as the selected group member may not well specify the population as a whole.

The unbiased atlas represents an average anatomical configuration of a population. Unbiased atlas construction is an active area of research in image registration. Thompson and Toga [97] very elegantly address the bias problem by mapping a new data set onto every scan in a brain image database. This approach addresses bias by forgoing the formal construction of a representative template image. Although this framework is mathematically elegant and powerful, it results in a computationally prohibitive approach in which each new scan has to be mapped independently to all datasets in a database. This is analogous to comparing each subject under study to every previously analyzed image. As brain image databases grow, the atlas formation problem grows combinatorially.

Previous work in atlas formation has focused on the small deformation setting in which arithmetic averaging of displacement fields is well defined, e.g., in [9]. Studholme minimizes an energy functional involving uncertainty in joint histograms of intensities, elastic deformation, and sum of displacements [90]. Kovacevic et al. present a multiresolution method that is initialized by averaging pairwise affine transformations and finalized by centering based on averaging estimated non-linear transformations to this affine average [57]. An iterative averaging algorithm to reduce the bias has been developed by [34]. In the latest work of [9], explicit constraints requiring that the sum of the displacement fields add to zero are enforced in the proposed atlas construction methodology. These small deformation approaches are based on the assumption that transformations of the form $h(x) = x + u(x)$, parameterized via a displacement field, $u(x)$, are close enough to the identity transformation such that the composition of any

two transformations can be approximated via the addition of their displacement fields:

$$(h_1 \circ h_2)(x) \approx x + u_1(x) + u_2(x).$$

Using a hidden probabilistic model of the common spatial distribution of anatomical tissues, De Craene et al. create atlases of probability distributions using STAPLE [17]. The Simultaneous Truth and Performance Level Estimation (STAPLE) method, developed by Warfield and Zou [104, 112], calculates a composite gold standard estimate from multiple manual segmentations. Given a set of binary segmentations of the same object, STAPLE calculates the maximum likelihood estimate of the composite "gold standard" or the best estimate of the unknown gold standard. The STAPLE algorithm calculates the specificity and sensitivity of each segmentation in an iterative way. In the work of De Craene et al., a generalized expectation maximization method is used where, in the expection step, atlas labels (hidden data) are estimated given fixed transformations and, in the in the maximization step, transformations that maximize a similarity criterion are estimated.

In a more recent and related work, Avants and Gee [5] develop an algorithm in the large deformation diffeomorphic setting by averaging velocity fields and evolving mean geodesic flows. The focus of this chapter is on the development of a methodology that simultaneously estimates the transformations and an unbiased template, in the large deformation setting. This method does not assume the above approximation and, thus, is capable of building atlases of populations with large geometric variability. The method proposed in this dissertation is intrinsically unbiased in that it involves no penalty terms in the optimization process. The method is also computationally efficient in that it scales linearly with the number of images. Before formally defining the atlas formation problem, this chapter explores averaging large deformation diffeomorphic transformations.

4.2 Averaging Diffeomorphisms

Given a collection of anatomical images, a natural problem is the construction of a statistical representative of the population. If the data associated with the population under study can be easily parameterized by a Euclidean space, classical statistical methods of simple averaging can be applied to generate such a representative. An image under the Gaussian noise assumption can itself be easily represented as a member

of a flat space. The image can be represented as a member of a very large dimensional Euclidean space $\mathbb{R}^N$, where N is the number of voxels in the image. Alternatively, using appropriate interpolation assumptions the image can be assumed to be a square integrable function, that is a member of the Hilbert space $L^2(\Omega)$ where Ω is the underlying coordinate space, usually a compact subset of $\mathbb{R}^3$.

The geometric variability of the anatomy itself usually cannot be represented by elements of a flat space. If the geometry of the underlying anatomy can be adequately represented by a finite number of landmarks, representative template landmark configuration can be estimated using the Procrustes method pioneered by Kendal [54] and championed by Bookstein [10]. The study of anatomical shape is inherently related to the construction of transformations of the underlying coordinate space that map one anatomy to another. Various transformation groups of $\mathbb{R}^3$ have been studied for understanding anatomical geometry. These groups vary in dimensionality from simple global translations, $\mathbb{R}^3$, and rigid rotations, $SO(3)$, to the infinite dimensional group of diffeomorphisms, $\mathcal{H}$, [71].

In this chapter, the problem of building an anatomical template is posed as a statistical estimation problem. For anatomical representations in which the underlying geometry is parameterized as a Euclidean vector space, training data can be represented as a set of vectors, e.g., $\{x_i\}_{i=1}^N$ in a vector space V.

In the small deformation elastic image mapping setting, this is assumed to be true, as the deformations are assumed to be close enough to the identity mapping. Under this assumption, the displacement vector fields parameterizing the transformations can be assumed to be elements of the Hilbert space of square integrable functions $L^2(\Omega)$.

In a vector space, with addition and scalar multiplication well defined, the average representation of the training set is the linear average

$$\bar{x}_{Linear} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{4.1}$$

In terms of computing statistics, the group of diffeomorphic transformations presents a challenge. Linear averaging cannot be directly applied to the large deformation setting, as under the large deformation model the space of transformations is not a vector space but rather the infinite dimensional group $\mathcal{H}$ of diffeomorphisms of the underlying domain Ω.

In the group of diffeomorphisms, the addition of two diffeomorphisms is not generally

a diffeomorphism and hence a template based on linear averaging of transformations is not well defined. To address this difficulty, the notion of averaging is extended to general metric spaces first proposed by Fréchet [27]. For a general metric space $\mathcal{M}$, with a distance $d : \mathcal{M} \times \mathcal{M} \rightarrow \mathbb{R}$, the *intrinsic mean* for a collection of data points x_i can be defined as the minimizer of the sum-of-squared distances to each of the data points. That is

$$\bar{x}_{Frechet} = \underset{x \in \mathcal{M}}{\operatorname{argmin}} \sum_{i=1}^{N} d(x, x_i)^2. \tag{4.2}$$

In previous work at UNC Chapel Hill, these concepts have been used to extend first and second order statistical analysis to finite dimensional Riemannian Manifolds for statistical analysis of medial representations of objects [26]. This chapter applies this approach to the construction of large deformation diffeomorphic templates. The work here builds heavily on the mathematical metric theory of diffeomorphisms developed by Miller and Younes [68].

Given a metric on a group of transformations, the atlas construction problem can be stated: estimate a class probability $\hat{p}$ that requires the minimum amount of deformation energy to be transformed into every member of the population of class posteriors, p_i. More precisely, given a transformation group $\mathcal{S}$ with associated metric $D : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$, along with a probability dissimilarity metric $E(p_1, p_2)$, to find the probability $\hat{p}$ such that

$$\{\hat{h}_i, \hat{p}\} = \underset{h_i \in \mathcal{S}, p}{\operatorname{argmin}} \left[\sum_{i=1}^{N} E(p_i \circ h_i, p) + D(e, h_i)^2 \right] \tag{4.3}$$

where e is the identity transformation and $D(e, h)$ is measure of how far from the identity transformation h is as descibed in Chapter 3.

This chapter addresses the problem of anatomical template construction as the joint the estimation of the most representative, average, image and, as encoded in transformations, the associated anatomical geometry given a database of brain images.

4.3 Large Deformation Diffeomorphic Atlas Estimation

Given the generalized notion of averages for diffeomorphic transformations, the class-conditional atlas estimation is defined as the probability mass function that minimizes an image dissimilarity measure and requires the least amount of energy, based

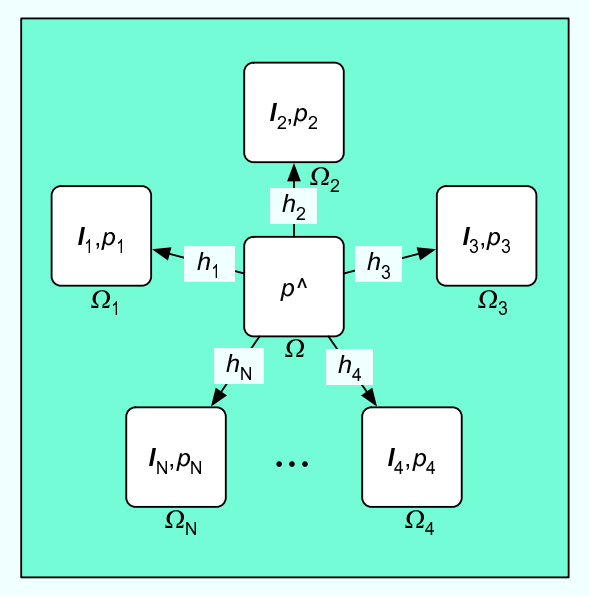


Figure 4.1: Atlas Formation

Unbiased atlas constructed as the intrinsic mean of a population of class posteriors.

on a metric on diffeomorphisms, to deform into the each member of the population. Specifically, consider the problem of estimating an atlas class posterior $\hat{p}$ that is the best representative for a population of N class posteriors, $\{p_i\}_{i=1}^N$, representing the N individual image sets $\{\bar{I}_i\}_{i=1}^N$. The atlas $\hat{p}$ is not a member of the set $\{p_i\}$. To this end, consider the problem of constructing a mapping between $\hat{p}$ and each class posterior in the set $\{p_i\}$. That is, estimate the mappings $h_i : \Omega \rightarrow \Omega_i$ where $\Omega \subset \mathbb{R}^3$ and $\Omega_i \subset \mathbb{R}^3$ are the domains of the class posteriors $\hat{p}$ and p_i respectively. The domain Ω is chosen to be independent of the individual population class posterior domains, Ω_i. This framework is depicted in Figure 4.1.

Using the metric on the space of diffeomorphisms with regularization operator L defined in Equation 3.5, the minimum energy atlas estimation problem expressed in Equation 4.3 is formulated as

$$\{\hat{h}_i, \hat{p}\} \quad = \quad \underset{h_i, p}{\operatorname{argmin}} \left[\sum_{i=1}^{N} E(p_i \circ h_i, p) + \int_0^1 \int_\Omega ||Lv_i(x,t)||^2 dx dt \right] \tag{4.4}$$

subject to

$$h_i(x) \;=\; x + \int_0^1 v_i(h_i(x,t),t)dt.$$

Note that the solution to this minimization problem is independent of the ordering of the N images. This atlas construction framework produces transformations $\hat{h}_i$ such that $\hat{h}_i : \Omega \to \Omega_i$. Since each $\hat{h}_i$ is a diffeomorphism, its inverse $\hat{h}_i^{-1} : \Omega_i \to \Omega$ exists and can be calculated by integrating the negative velocity fields backwards in time, see Figure 3.6. Image to image correspondences can be computed from these transformations using the composition rule

$$\hat{h}_{i,j} = \hat{h}_j \circ \hat{h}_i^{-1} : \Omega_i \to \Omega_j. \tag{4.5}$$

4.3.1 Dispersion Functions on the Space of Probability Mass Functions

The $\text{argmin}_p \sum_{i=1}^N E(p_i \circ h_i, p)$ term in Equation 4.4 represents the dispersion between the class-posteriors $\{p_i \circ h_i\}_{i=1\ldots N}$. Consider three such dispersion terms on the space of probability mass functions. Two of these are information distances presented in Chapter 2 and one is an extension of a well-known metric.

1. *Sum-of-Kullback-Leibler Divergences I*

$$\bar{D}\left(\hat{p}(x)||\{p_i(h_i(x))\}_{i=1}^N\right) = \sum_{i=1}^N D(\hat{p}(x)||p_i(h_i(x)))$$

   where $\hat{p}$ is the normalized geometric mean of $\{p_i\}_{i=1}^N$.

2. *Sum-of-Kullback-Leibler Divergences II*

$$\bar{D}\left(\{p_i(h_i(x))\}_{i=1}^N||\hat{p}(x)\right) = \sum_{j=1}^N D(p_i(h_i(x))||\hat{p}(x))$$

   where $\hat{p}$ is the arithmetic mean of $\{p_i\}_{i=1}^N$

3. *Sum-over-Class Squared Error*

$$D_{SE}(\{p_i(h_i(x))\}_{i=1}^N) = \sum_{c\in C}\sum_{i=1}^{N}\left(p_i(h_i(c(x))) - p(c(x))\right)^2$$

The first two dispersions are sum-of-Kullback-Leibler divergences defined in Chapter 2. Minimizing either $\bar{D}\left(\hat{p}(x)||\{p_i(h_i(x))\}_{i=1}^N\right)$ or $\bar{D}\left(\{p_i(h_i(x))\}_{i=1}^N||\hat{p}(x)\right)$ between $\{p_i\}_{i=1}^N$, maximizes a lower bound on Bayes' probability of error P_e and thus renders the probability mass functions more indistinguishable. That is, it brings them closer together. The third dispersion is the extension of the squared error dissimilarity measure used for scalar images used in [49] to probability mass functions. Although the author has not found a relationship between D_{SE} and P_e, one can see that when $D_{SE} = 0$, $P_e = \frac{N-1}{N}$ and is maximal. Moreover, D_{SE} is a true metric.

4.3.2 Registration

With these results, the minimization problem stated in Equation 4.4 can be further specified in one of three ways corresponding to the choice of distance function.

1. *Sum-of-Kullback-Leibler Divergence I*

$$\hat{v}_i = \underset{v_i}{\operatorname{argmin}}\left[\int_\Omega \sum_{i=1}^{N} D(p(x)||p_i(h_i(x)))dx + \int_0^1\int_\Omega ||Lv_i(x,t)||^2 dxdt\right] \tag{4.6}$$

   subject to $h_i(x) = x + \int_0^1 v_i(h_i(x,t))dt$.

2. *Sum-of-Kullback-Leibler Divergence II*

$$\hat{v}_i = \underset{v_i}{\operatorname{argmin}}\left[\int_\Omega \sum_{j=1}^{N} D(p_i(h_i(x))||p(x))dx + \int_0^1\int_\Omega ||Lv_i(x,t)||^2 dxdt\right] \tag{4.7}$$

   subject to $h_i(x) = x + \int_0^1 v_i(h_i(x,t))dt$.

3. *Sum-over-Class Squared Error*

$$\hat{v}_i = \int_\Omega \sum_{c\in C}\sum_{i=1}^{N}\left(p_i(h_i(c(x))) - p(c(x))\right)^2 dx + \int_0^1\int_\Omega ||Lv_i(x,t)||^2 dxdt \tag{4.8}$$

   subject to $h_i(x) = x + \int_0^1 v_i(h_i(x,t))dt$.

Note that the solution to any of these three minimization problem is independent of the ordering of the N image sets and increases linearly as image sets are added, thus, making the algorithm scalable.

4.3.3 Variation of Dispersion With Respect to Transformations

A gradient descent approach to optimizing the estimation problems in Section 4.3.2 is used. To that end, the variation of the dispersion function with respect to the transformation h_i can be computed as follows:

1. *Sum-of-Kullback-Leibler Divergences I*

$$\begin{aligned}\frac{\partial}{\partial h_i}\bar{D}\left(\hat{p}(x)||\{p_j(h_j(x))\}_{i=1}^N\right) &= \frac{\partial}{\partial h_i}\sum_{j=1}^N D(p(x)||p_j(h_j(x))) \\ &= \frac{\partial}{\partial h_i}\sum_{c\in C} p(c(x))\log\frac{p(c(x))}{p_i(h_i(c(x)))} \\ &= -\sum_{c\in C}\frac{p(c(x))}{p_i(h_i(c(x)))}\nabla p_i|_{h_i(c(x))}^T.\end{aligned}$$

2. *Sum-of-Kullback-Leibler Divergences II*

$$\begin{aligned}\frac{\partial}{\partial h_i}\bar{D}\left(\{p_j(h_j(x))\}_{j=1}^N||\hat{p}(x)\right) &= \frac{\partial}{\partial h_i}\sum_{j=1}^N D(p_j(h_j(x))||p(x)) \\ &= \frac{\partial}{\partial h_i}\sum_{c\in C} p_i(h_i(c(x)))\log\frac{p_i(h_i(c(x)))}{p(c(x))} \\ &= \sum_{c\in C}\left[\log\frac{p_i(h_i(c(x)))}{p(c(x))}+1\right]\nabla p_i|_{h_i(c(x))}^T.\end{aligned}$$

3. *Sum-over-Class Squared Error*

$$\begin{aligned}\frac{\partial}{\partial h_i}D_{SD}(\{p_j(h_j(x))\}_{j=1}^N) &= \frac{\partial}{\partial h_i}\sum_{c\in C}\sum_{j=1}^N\left(p_j(h_j(c(x)))-p(c(x))\right)^2 \\ &= \sum_{c\in C}\frac{\partial}{\partial h_i}\left(p_i(h_i(c(x)))-p(c(x))\right)^2 \\ &= 2\sum_{c\in C}\left(p_i(h_i(c(x)))-p(c(x))\right)\nabla p_i|_{h_i(c(x))}^T.\end{aligned}$$

4.3.4 Implementation

Given one of the three minimization problems, the iterative greedy fluid algorithm of propagation templated described in [14] is used to approximate the solution. At each iteration k, the updated transformation h_i^{k+1}, for each class-conditional probability p_i, is computed using the update rule $h_i^{k+1} = h_i^k \left(x + \varepsilon v_i^k(x)\right)$. h_i^k and v_i^k are the current estimated transformation and velocity for the ith probability, and ε is the step size. In other words, each final transformation h_i is built up from the composition of k transformations.

The velocity v_i^k for each iteration k is computed as follows. First, compute the updated template estimate. For the optimization in Equation 4.6, this is normalized geometric mean,

$$\hat{p}^k(c(x)) = \frac{\left(\prod_{i=1}^N p_i(h_i^k(c(x)))\right)^{\frac{1}{N}}}{\sum_{c' \in C}\left(\prod_{i=1}^N p_i(h_i^k(c'(x)))\right)^{\frac{1}{N}}},$$

for each class component c. For the optimizations in Equations 4.7 and 4.8, this is the arithmetic mean,

$$\hat{p}^n(c(x)) = \frac{1}{N}\sum_{i=1}^N p_i(h_i^n(c(c))),$$

for each class component c. Next, define the body force functions as the variation of the class posterior dispersion terms with respect to the transformation h_i.

1. *Sum-of-Kullback-Leibler Divergences I*

$$F_i^k(x) = -\sum_{c \in C} \frac{p(c(x))}{p_i(h_i^k(c(x)))} \nabla p_i|_{h_i(c(x))}^T.$$

2. *Sum-of-Kullback-Leibler Divergences II*

$$F_i^k(x) = \sum_{c \in C}\left[\log \frac{p_i(h_i^k(c(x)))}{p(c(x))} + 1\right] \nabla p_i|_{h_i^k(c(x))}^T.$$

3. *Sum-over-Class Squared Error*

$$F_i^k(x) = 2\sum_{c \in C}\left(p_i(h_i^k(c(x))) - p(c(x))\right) \nabla p_i|_{h_i^k(c(x))}^T$$

The velocity fields are estimation and integrated to produce subsequent forward and inverse transformations as before. The velocity field v_i^k is computed at each iteration by applying the inverse of the differential operator L to the force function, i.e. $v_i^k(x) = L^{-1}F_i^k(x)$, where $L = \alpha\nabla^2 + \beta\nabla(\nabla\cdot) + \gamma$ is the Navier-Stokes operator. This computation is carried out in the Fourier domain [51].

For each iteration the dominating computation is the Fast Fourier Transform. Thus, the order of the algorithm is $MNn\log n$ where M is the number of iterations, N is the number of images to be registered, and n is the number of voxels in each image. The complexity increases only linearly as images are added, making the algorithm extremely scalable. Satisfactory correspondence is typically achieved after 100-200 iterations. In practice, a multi scale approach that initializes the fine (voxel) scale registration is used with the up-sampled correspondence computed at a coarser scale level. The finer scale levels only need to account for residue from coarser scale levels and thus require far fewer iterations to converge.

4.4 Affine and Diffeomorphic Atlas Results

To evaluate the performance of the atlas formation method, the algorithm, with sum-over-class squared error distance in Equation 4.8, was applied to a set of ten class-posterior mass function maps the database of healthy normal adult brains described in Chapter 6. A mid-axial slice from each class-posterior is shown in Figure 4.2. There is noticeable variation between these anatomies, especially in the ventricular region.

Figure 4.3 shows the arithmetic mean of the class posterior population following affine alignment and the final large deformation diffeomorphic average atlas estimate. The arithmetic mean is blurry since it is an "average" of the varying individual neuroanatomies. Ghosting is evident around the lateral ventricles and near the boundary of the brain. In the final estimate of non-linear atlas, these variations have been accommodated by the high-dimensional diffeomorphic registration.

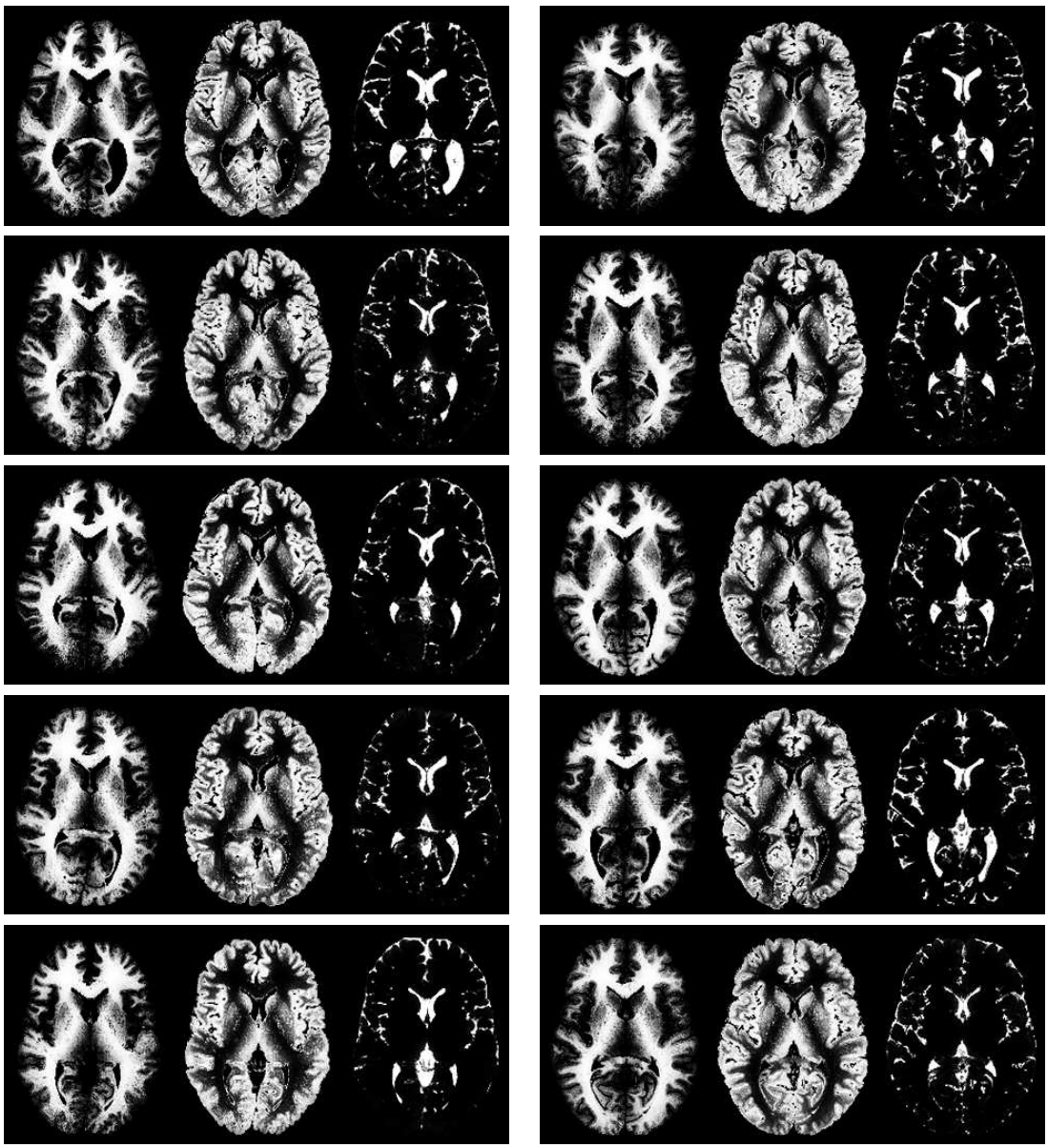
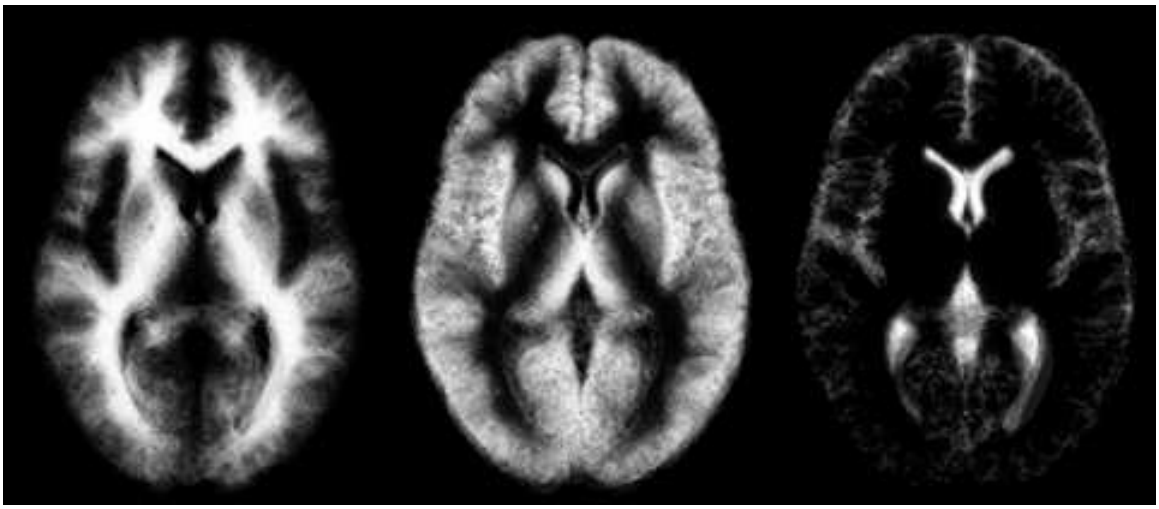
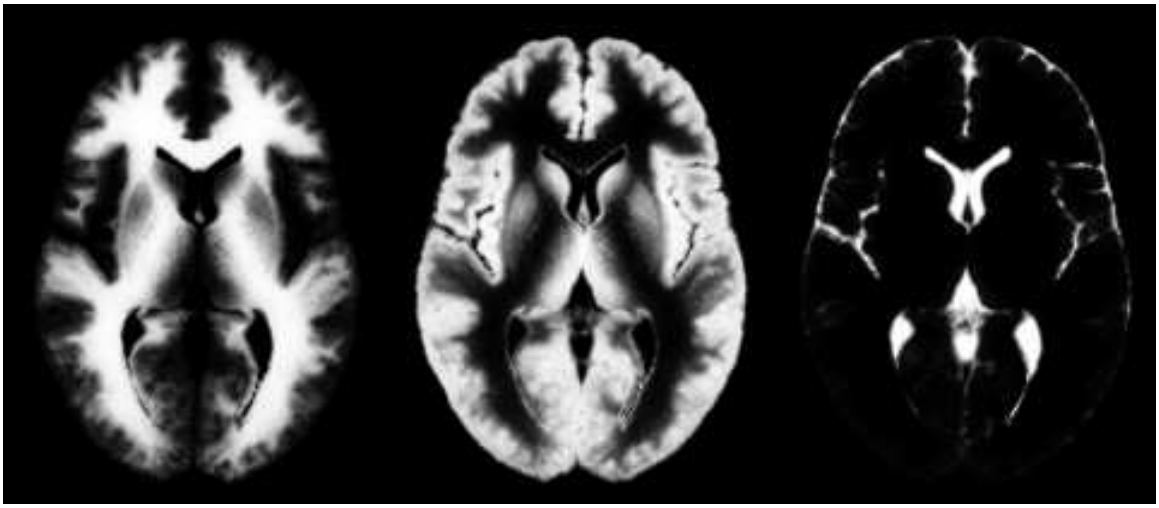


Figure 4.2: Sample Class Posterior Population

Axial views of ten class-posteriors. These images clearly indicate large inter-subject variability, especially in the ventricular system.



(a) Affine Average



(b) Large Deformation Diffeomorphic Average

Figure 4.3: Affine and Diffeomorphic Atlases

Axial views of the simple linear averaging of ten class-posteriors (a) and large deformation diffeomorphic averaging of the same ten class-posteriors. The diffeomorphic averaging produces a sharper atlas.

Chapter 5

Atlas Stability[1]

An important question to ask of the atlas formation method in Chapter 4 is the following: how many subjects are required to represent a population? That is, how many images are necessary to form a stable atlas? For a given population of images and corresponding atlas, there are a number of ways to determine if an additional image would result in a sufficiently different atlas. In this dissertation, entropy and variance of scalar fields, e.g. image intensities in T1-weighted MR images and the white matter class in a multi-class posterior map, is used to assess the stability of atlases. This chapter focuses primarily on the entropy of scalar intensity images. For atlas stability, it is the entropy introduced by the atlas creation method that is of interest rather than the intrinsic entropy associated with images of individual brain anatomy.

Entropy, defined in Appendix A, has often been proposed as a good measure of image quality [33, 4] where sharp images have relatively low entropy. A distribution $p_{single}(i) = \{1\}$ for a single event random variable, e.g. an image with constant intensity, has minimal entropy, $H(p_{single}) = 0$. The uniform distribution $p_{unif}(i) = \left\{\frac{1}{N}, \cdots, \frac{1}{N}\right\}$, for a random variable with at least two events, corresponding to an image with multiple intensities occurring with equal frequency, maximizes entropy, $H(p_{unif}) = \log N$ [16]. That is, a blurry image, with a relatively flat histogram, will have greater entropy than a sharp image. Using multiple permutations of images from a database of images, the stability of atlases produced by proposed atlas formation method is studied by building atlases of increasing population size. An atlas is considered stable when the entropy of its intensities or class-posterior maps is stable with respective to the number of constituent images.

The remainder of this chapter is organized as follows. For illustrative purposes, the

[1] This chapter is an extension of portions of the recent MICCAI paper [61].

entropy computation for a simple intensity image is provided in Section 5.1. Image intensity entropy is explored in terms of interpolation and scale in Section 5.2. In Section 5.3, a random permutation test is used to study atlas stability with respect to the number of images used in the atlas formation. The chapter concludes with a summary in Section 5.4.

5.1 Simple Intensity Example

Consider the square 2D intensity image involving twenty-five voxels and associated height field interpretation shown in Figure 5.1. As a random variable, this image has five events corresponding to the image intensities, $E = \{50, 100, 150, 200, 250\}$. The probability mass function associated with this random variable is computed as the normalized histogram of the image intensities. For the intensities of this simple image, the normalized histogram is shown in Figure 5.2. The entropy of image intensities is computed from the probability mass function $p = \{0.48, 0.32, 0.04, 0.08, 0.08\}$ as

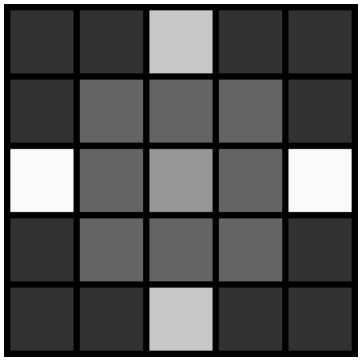
$$\begin{aligned} H(p) &= -\sum_{i=1}^{5} p(i) \log p(i) \\ &\approx 1.8 \text{ bits/event.} \end{aligned}$$

That is, on average, the minimum number of bits to represent an intensity value in this simple image is $\lceil H(p) \rceil = 2$.

5.2 Entropy: Interpolation and Scale Effects

This section focuses on quantifying the effects on entropy of both the choice of interpolation method for image resampling and viewing images through various apertures or at scale. In this, and subsequent results in this chapter, 256 histogram bins were used to define the probability mass functions. Therefore, $\log_2 256 = 8$ bits is the upper limit for the entropy computation in these experiments.

To examine the effects of interpolation on entropy, a simple binary image was constructed and translated over a range of distances using four different interpolation methods during the resampling. After each resampling, the entropy of image intensities was computed. The simple image was chosen to be a uniformly white disk on a uniformly black background. To reduce bias in the entropy computation, the areas of



50	**50**	**250**	**50**	**50**
50	**100**	**100**	**100**	**50**
200	**100**	**150**	**100**	**200**
50	**100**	**100**	**100**	**50**
50	**50**	**250**	**50**	**50**

Figure 5.1: Simple Intensity Image

A simple 5×5 image with five distinct intensities and corresponding scalar height field interpretation are shown on the left and right respectively.

Event (Intensity)	**Count**	**Frequency**
50	12	0.48
100	8	0.32
150	1	0.04
200	2	0.08
250	2	0.08

Figure 5.2: Image Intensity Histogram

The five intensities in the image, the associated histogram, and normalized (to unity) histogram are shown in the left, middle, and right columns respectively.

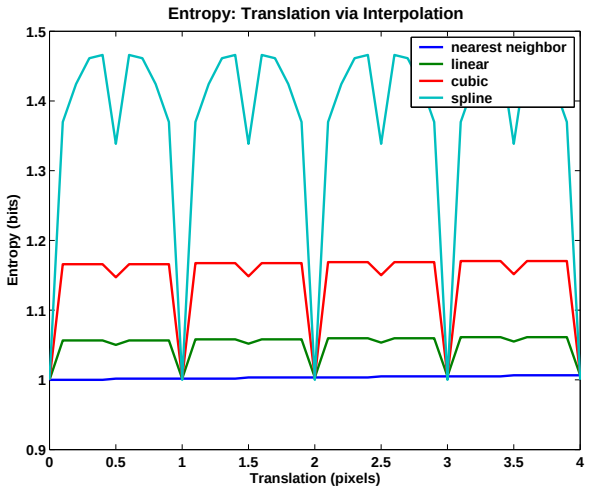


Figure 5.3: Translated Disk Entropy

Entropy plots for the disk image translated over a range of forty distances, uniformly spaced over four voxels, using four interpolation methods: nearest neighbor, linear, cubic, and spline.

the disk and background were constructed to be equal. The results of translating this disk are shown in Figure 5.3. The higher order interpolation methods induce greater entropy. Any evaluation of entropy for image quality will, therefore, have to consider the method of interpolation used during image resampling.

To examine the effects on entropy of viewing an image at scale, the same binary disk image was low-pass filtered with fifteen Gaussian kernels of increasing scale. The original image and the fifteen blurred versions thereof are shown in Figure 5.4. The entropy of the original and blurred disks is presented in Figure 5.5. For this example, entropy grows linearly with scale σ. The slope of the line is approximately 0.2 bits/σ. The entropy of the base image is appropriately one bit. That is, a pixel is either white or black in equal proportions.

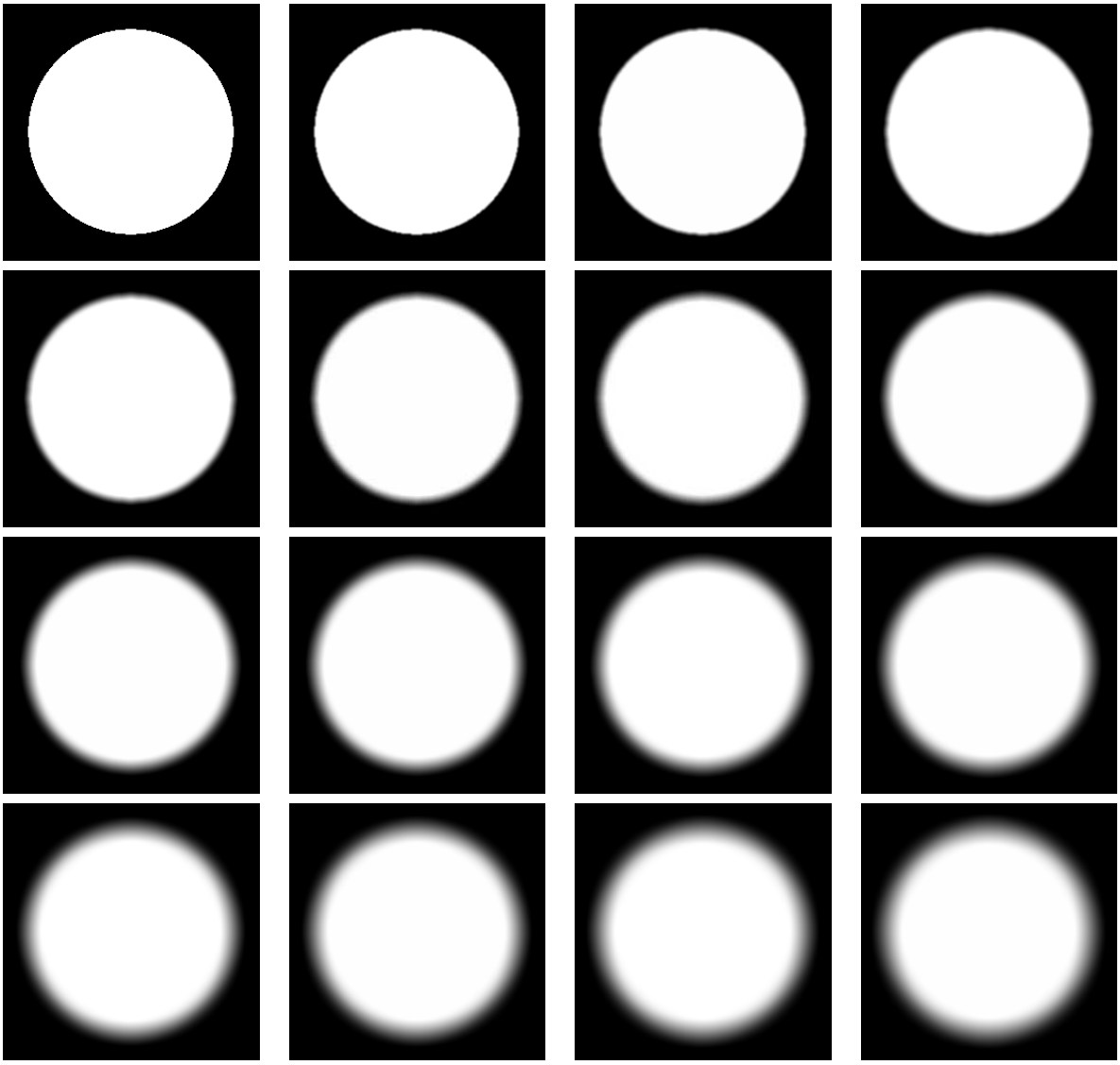


Figure 5.4: Disk at Scale

The original disk, 256×256 voxels, (upper right) and the fifteen blurred versions produced by convolution with a Gaussian filter with scale parameters $\sigma = 1 \ldots 15$ pixels. The kernel width was defined to be $2\sigma + 1$ pixels wide.

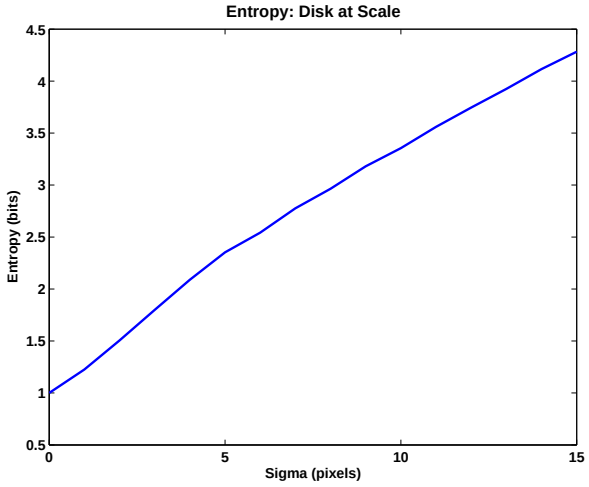


Figure 5.5: Disk at Scale Entropy

5.3 Random Permutation Test

To address the question of how many images are required to construct a stable atlas random permutation tests involving 2D scalar T1-weighted images were investigated. In that work [61], rigid-based and diffeomorphic-based atlases were compared in terms of stability through entropy of image intensities and variance in image intensities. The results from that prior work are reproduced in this section.

5.3.1 Atlas Formation

Atlases were created from an image database containing fourteen brain images provided by the UNC Chapel Hill autism image analysis group. These images were intensity normalized and rigidly aligned. Due to the high memory demands of the implementation, the algorithm was applied to 2D mid-axial slices. This database of images is shown in Figure 5.6. There is noticeable large deformation variation between these anatomies, especially in lateral ventricles.

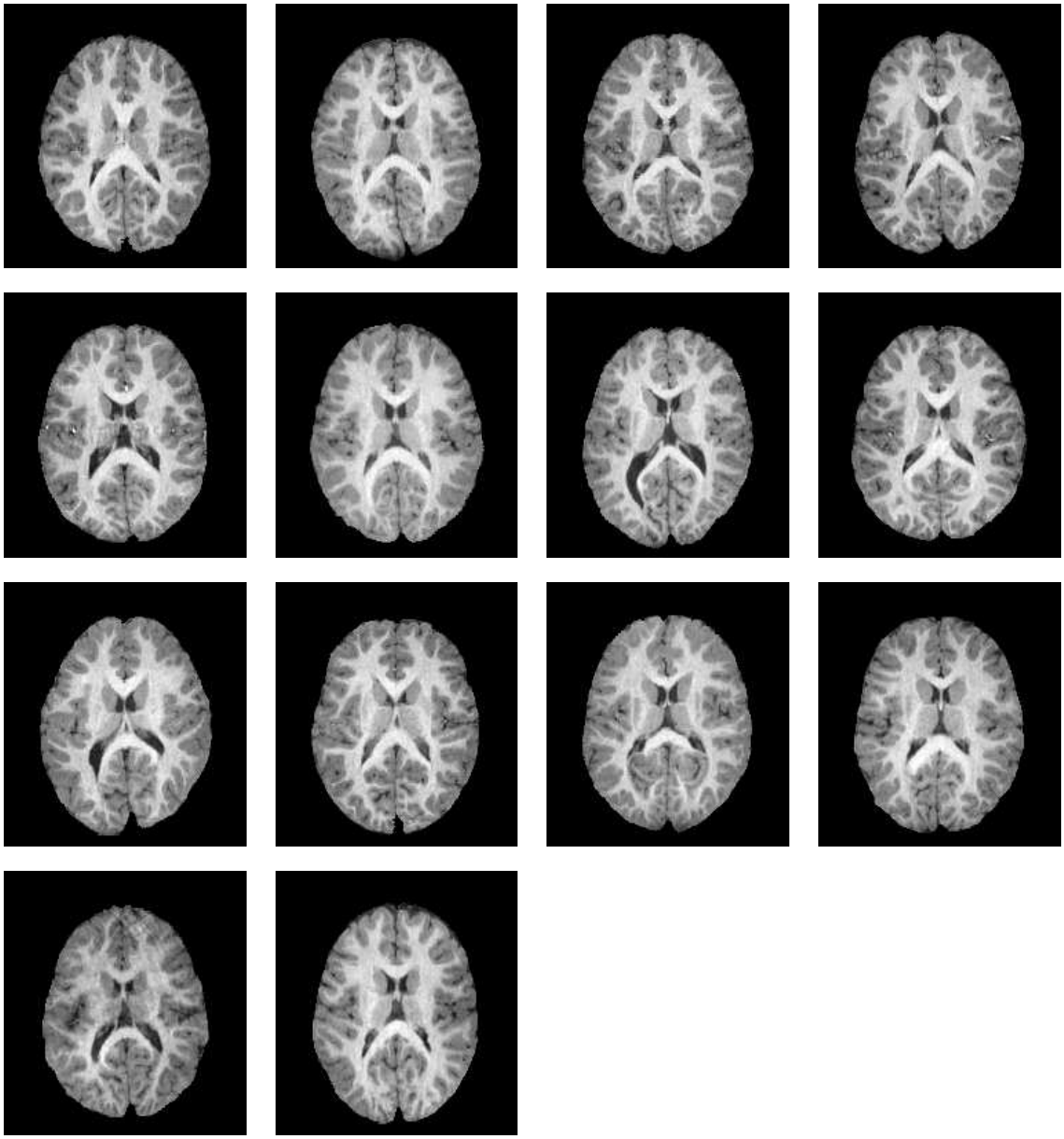


Figure 5.6: 2D Population Data

2D mid-axial slices from MR images of fourteen subjects.

To quantify the stability of the estimated atlases, eleven atlas cohorts, $\{C_i\}_{i=2}^{12}$, were generated each with twenty atlases derived from i images randomly selected from the original database of fourteen images. Two mutually exclusive atlases from C_7 are shown in Figure 5.7 for both simple averaging and with the large deformation diffeomorphisms. The rigidly aligned atlases are blurry since they are arithmetic averages of varying individual neuroanatomies. Ghosting is evident around the lateral ventricles and near the boundary of the brain. In the final diffeomorphic atlases, these regions appear much sharper.

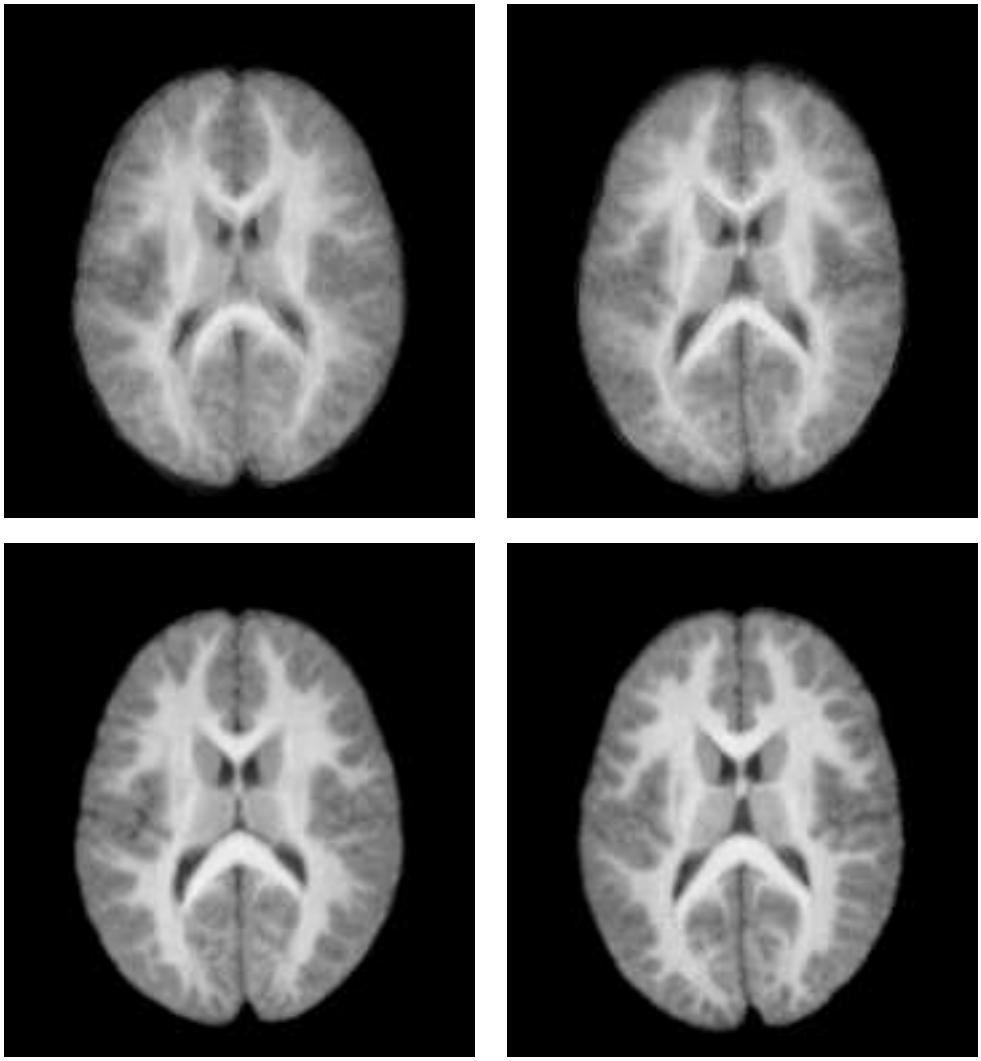


Figure 5.7: Mutually Exclusive Atlases

Both columns represents an individual atlas constructed by both arithmetically averaging rigidly aligned images (top row) and estimating a diffeomorphic atlas after 100 iterations (bottom row). These two atlases were formed from completely separate sets of images.

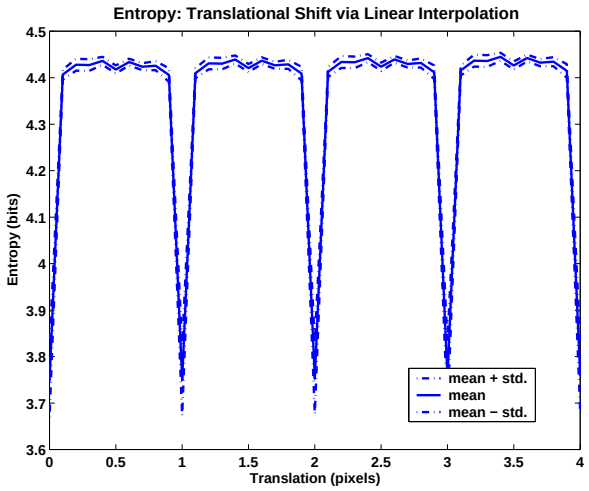


Figure 5.8: Translated T1 Image Entropy

Entropy plots for the fourteen images in the database over a range of forty distances, uniformly spaced over four voxels, using linear interpolation.

5.3.2 Interpolation and Scale Effects on Entropy

As in Section 5.2, the effects on entropy of both interpolation, as used in image resampling, and viewing images at scale are considered.

To examine the effects of interpolation on entropy, each of the fourteen images in the database were a translated over a range of distances using linear interpolation. After each resampling, the entropy of image intensities was computed. The mean and standard deviation of entropy of these fourteen images over the range of distances is shown in Figure 5.8. The fourteen images were intensity normalized and from subjects of similar age which may explain the rather tight variance observed in the entropy computation.

To examine the effects on entropy of viewing an image at scale, a single image taken from the database was low-pass filtered with fifteen Gaussian kernels of increasing scale. The original image and the fifteen blurred versions thereof are shown in Figure 5.9. The entropy of the original and blurred images is presented in Figure 5.10. As with the disk example, entropy grows linearly with scale σ, but after $\sigma = 1$ voxel. The slope of the line is approximately 0.1 bits/σ, half that of the disk example. It is interesting to note that, for this image, using linear interpolation increases the entropy to approximately 4.43 bits of entropy which is similar to the 4.59 bits observed by blurring the image

with $\sigma = 1$ voxels.

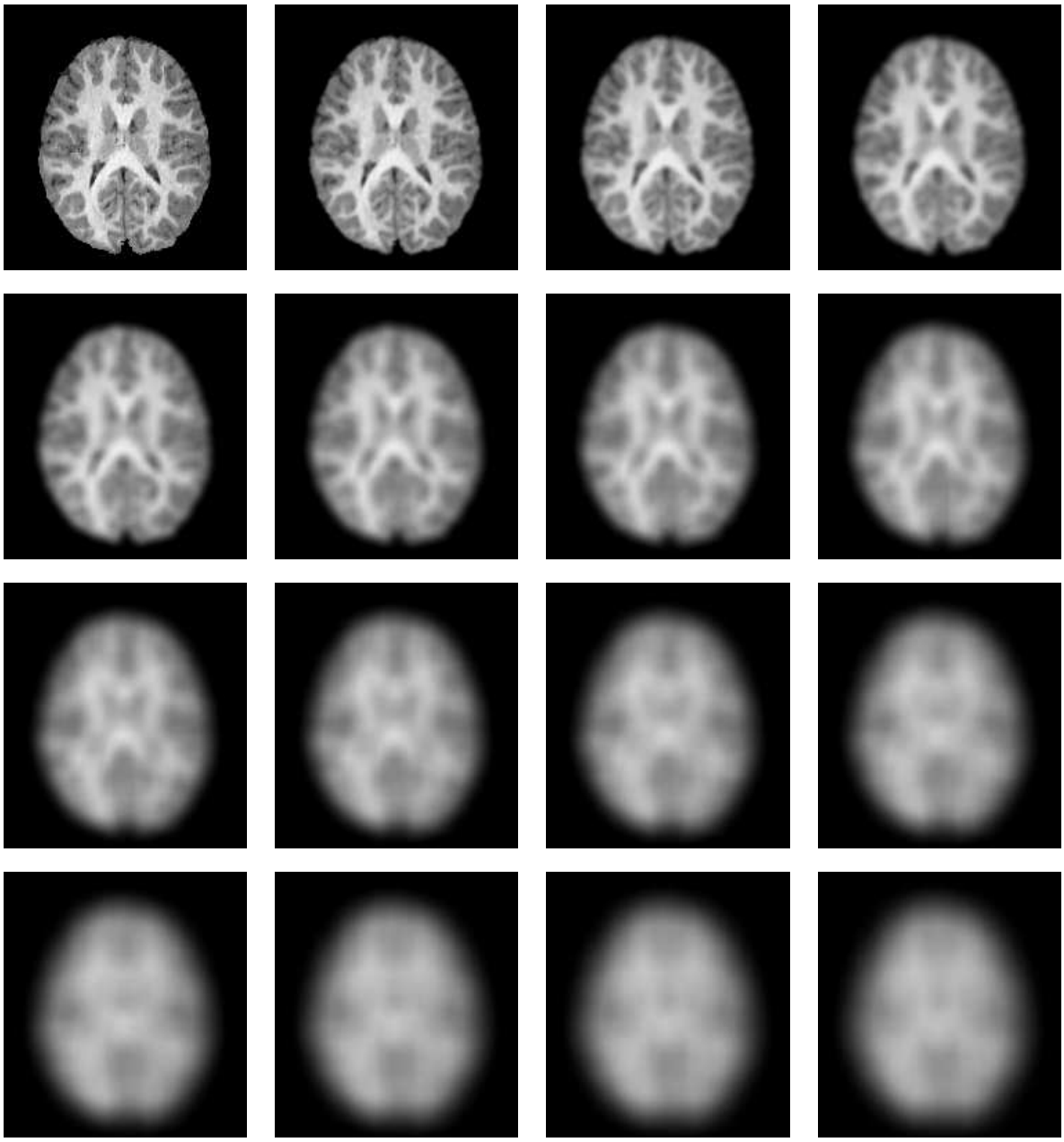


Figure 5.9: T1 Image at Scale

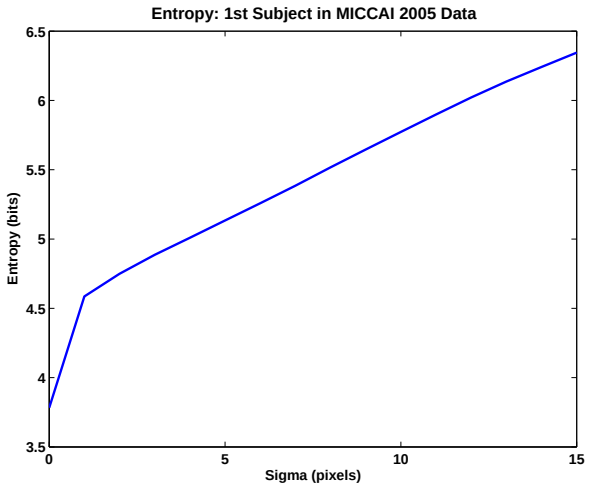


Figure 5.10: T1 Image at Scale Entropy

5.3.3 Atlas Stability

To evaluate the robustness and stability of the atlases, the mean and standard deviation of the entropies of the original fourteen images were first computed. To this end, the mean and standard deviation of the atlas cohort entropies created both by simple arithmetic averaging of the rigidly aligned images and those produced by the diffeomorphic method were compared. These results are summarized in Figure 5.11. From this plot, one can notice that as atlas size increases, the average atlas entropy increases for atlases formed by simple intensity averaging, whereas the average entropy decreases for atlases created via diffeomorphism. The atlases also become more stable with respect to entropy as the standard deviation decreases with atlas size. After cohort C_{10}, the atlas entropy means appear to converge. Note that the entropy of the large deformation diffeomorphic atlases converges to about 4.5 bits, approximately same entropy of any of the shifted original images in the database. That is, to within the choice of interpolation method, the large deformation diffeomorphic atlases are sharp.

Another measure of atlas stability is the variance in image intensities. For each atlas cohort, point-wise intensity mean and variance images were created. For a given cohort C_s representing the $N = 20$ atlases $\{A_i^s(x)\}_{i=1}^N$ of size s, the mean and variance images were computed as follows

$$M_s(x) = \frac{1}{N}\sum_{i=1}^{N} A_i^s(x)$$

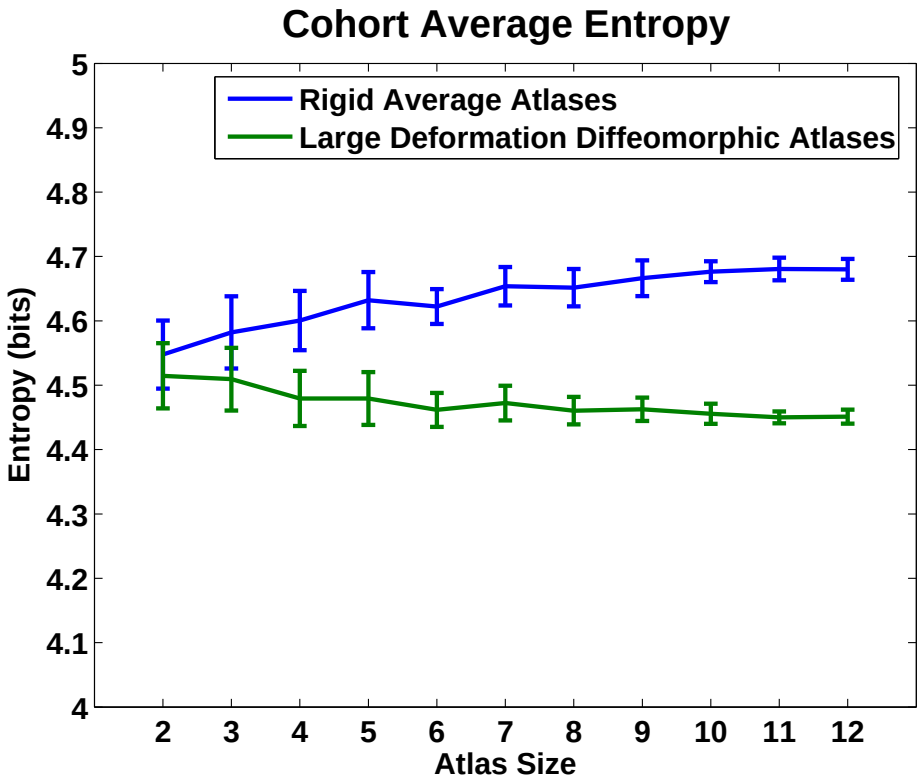


Figure 5.11: Cohort Average Entropy

For comparison, the average entropy of the original fourteen images is 3.91 bits with standard deviation 0.08 bits. The error bars represent one standard deviation from the mean.

and

$$V_s(x) = \frac{1}{N-1} \sum_{i=1}^{N} \left(A_i^s(x) - M_s(x)\right)^2 .$$

Cohort variance images for the large deformation diffeomorphic average and rigid average atlases are shown in Figures 5.12 and 5.13 respectively. These images show reduction of image intensity variation with the increase in the number of subjects per atlas. This reduction can be quantitatively observed through the relative variance sum $\frac{\mathcal{V}_s}{\mathcal{V}_2}$, where

$$\mathcal{V}_s = \sum_{x \in \Omega} V_s(x).$$

This measure was computed for both averaging methods and is shown in Figure 5.14.

Both methods show convergence in variance, which, as with the entropy measure, indicate atlas stability is achieved with about ten subjects for this dataset. That is, given these fourteen subjects, about ten images are needed to create a stable, with respect to the entropy and variance measures, atlas representing neuroanatomy.

5.4 Summary

In this chapter, entropy of image intensities was used to quantify atlas sharpness. A blurry image will exhibit greater entropy of intensities than a sharper image. Using this measure, it was shown that resampling an image using linear interpolation adds entropy to the image comparable to that of blurring the image with Gaussian kernel with a width $\sigma = 1$ voxels. For atlas formation, entropy and image intensity variance was used to address the question of how many subjects are required to produce a stable atlas. Random permutation tests were conducted to study atlas entropy and intensity variance as a function of the number of images used to produce an atlas. It was shown that, within the choice of interpolation method, the large deformation diffeomorphic atlases were sharp after the inclusion of ten or more subjects. This number is certainly dependent on the particular population used. Atlas stability for larger and more varied databases may require more subjects. The results in this chapter exemplify the atlas stability methodology rather than provide an answer for all image databases.

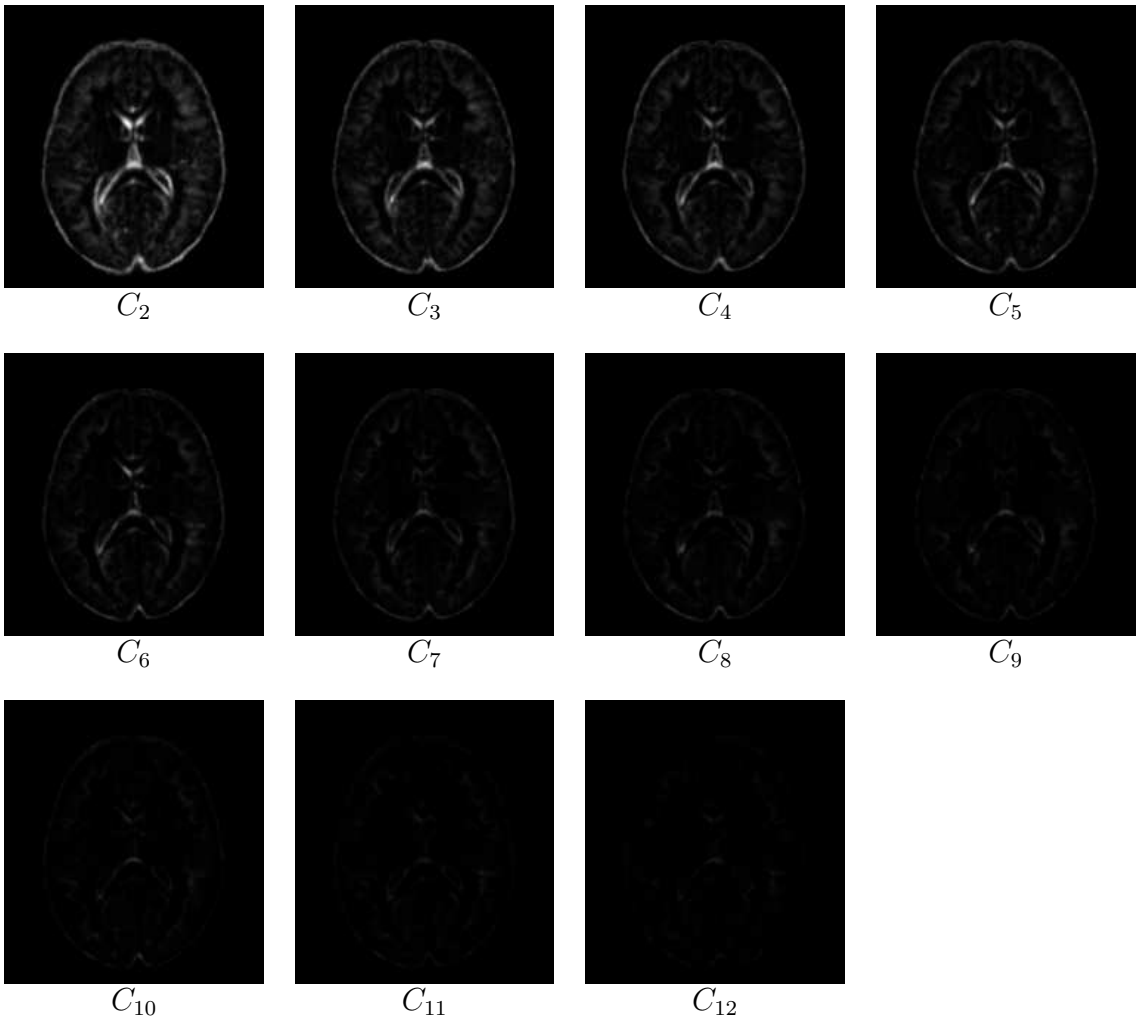


Figure 5.12: Cohort Variance: Large Deformation Diffeomorphic Atlases

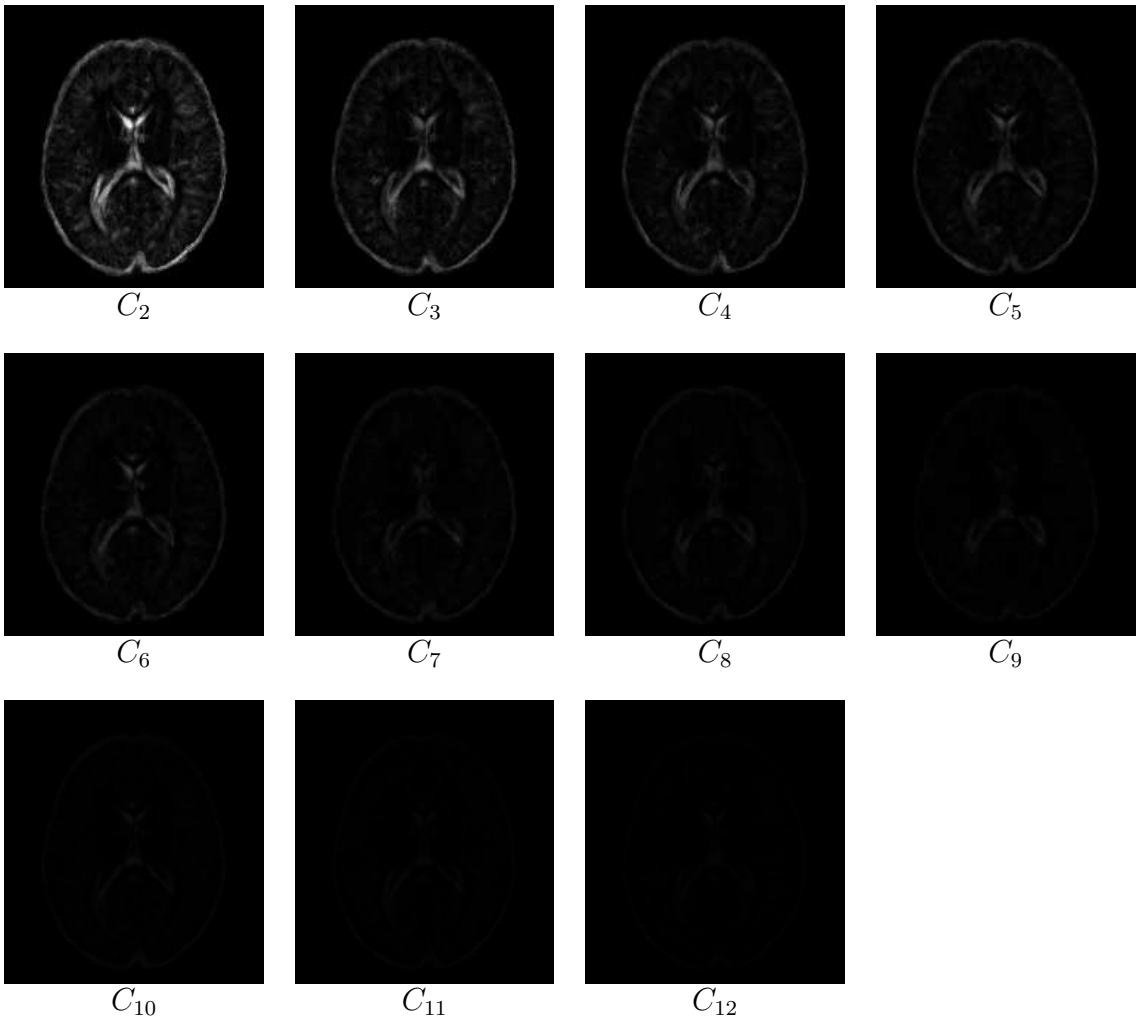


Figure 5.13: Cohort Variance: Rigid Average Atlases

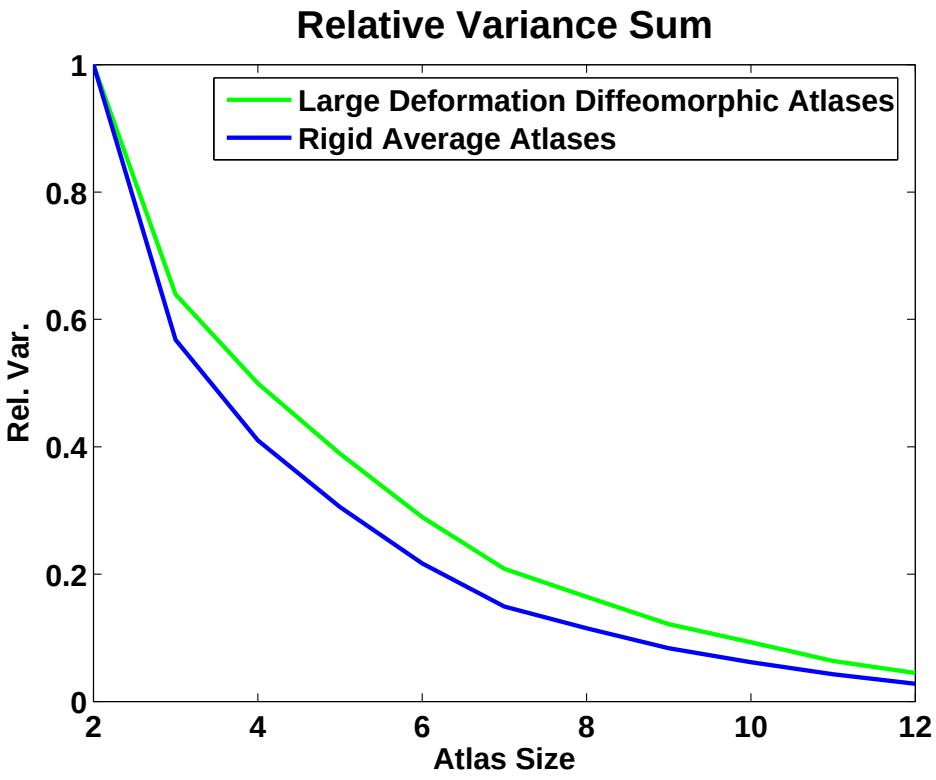


Figure 5.14: Cohort Variance: Relative Sum $\frac{\mathcal{V}_s}{\mathcal{V}_2}$

Chapter 6

Aging Study via Atlases of Normal Healthy Brains

An important area of medical image analysis is the development of methods for automated computer-assisted assessment of disease. For example, methods for analyzing local structural brain change over time can provide markers for understanding disease progression (e.g., Alzheimer's disease and schizophrenia). Early detection of disease-specific brain changes is important for therapeutic response. This requires the establishment of healthy norms to which test subjects can be compared. These norms typically take the form of a digital atlas constructed from data derived from many healthy subjects generally from a wide age range. Healthy aging, however, induces structural changes in the brain [67, 36]. In young, healthy adults [35] and in non-demented older persons [82], analysis demonstrates that the aging process is a continuum with changes evident before senescence. A comprehensive discussion of techniques used to examine the temporal dynamics of brain anatomy is provided in [100]. Before analyzing atlas-based brain change studies, the construction of brain databases is reviewed.

The design of brain databases can be categorized into two main types: longitudinal design and cross-sectional design. In longitudinal studies, the same subjects are scanned over time, typically with interscan intervals of one year. Cross-sectional studies, by contrast, involve scanning many subjects of different ages only once. While these studies capture brain changes of longer time periods, underlying brain changes are harder to detect. This is a result of intersubject variability. Multivariate modeling is often used in cross-sectional studies to partition variance observed in the database into specific effects such as age and sex. To account fully for these effects, the Computer-Assisted

Surgery and Imaging Laboratory (CASILab) at UNC Chapel Hill has constructed an age-grouped MR image database of healthy subjects [73]. A potentially diseased subject can then be appropriately age- and sex-matched to the image database.

Temporal brain change has been studied using a number of methods. An approach that models brain deformation in Alzheimer's disease is presented in [28]. This method uses longitudinal data with one-year intervals. Pair-wise, in time, 3D non-linear image registration was performed for nine control subjects and nine Alzheimer's patients. Warping images from one time point to the next provides some measure of temporal smoothing. Another longitudinal method that analyzes transformations between cortical maps of adolescent subjects using elastic image registration is described in [95].

A voxel-based morphometric analysis of white matter, grey matter, and cerebrospinal fluid volume change using a large, 465-adult-subject cross-sectional database is presented in [30]. This method involves mapping each subject's anatomy to a template followed by performing statistics on the resulting volumetric changes. The choice of a single reference template may introduce a bias. Although four times larger and covering a similar age range to the UNC database, this database is skewed to a young adult population in their twenties and thirties.

The above 3D methods may be sensitive to noisy longitudinal measurements. To address this issue, a fully 4D approach that simultaneously estimates longitudinal correspondence (intra-subject) and inter-subject correspondence between template and subject is presented in [89]. The method is a generalization of the 3D method that uses image intensity, edge, and geometric moment information as a feature vector proposed by the authors in [88]. A 4D match is obtained by mapping a set of "active" points. A "distinctiveness" measure is used to select these "active" points. During the registration, a hierarchical approach is taken beginning with a smaller number of more distinct points and progressing to a larger number of less distinct points. This method was applied to nine subjects taken from the Baltimore Longitudinal Study of Aging [82] where each subject was scanned once a year for five years.

An age-continuous 4D spatiotemporal atlas would address the need for sufficient temporal resolution. In this chapter, the multi-class posterior atlas formation method developed in Chapter 4 is used to produce time sequences by age of atlases, resulting in 4D spatiotemporal atlases. These discrete 4D atlases provide a step toward building age-continuous 4D spatiotemporal atlases. A continuous 4D spatiotemporal atlas would provide the ideal age-based match for a new subject. By examining the Jacobian maps of dense transformations relating the space of the first atlas in the sequence to the

space of each subsequent atlas, local volumetric change through time can be studied. For this purpose, a program called PMFAtlasBuilder was written using C++ libraries developed at the radiation oncology department at UNC Chapel Hill.

The rest of this chapter is organized as follows. In Section 6.1, the database of multi-modal MR brain images is detailed. The creation of the class-conditional posterior maps representing the anatomical structures in the MR database is described in Section 6.2. The construction of individual spatial atlases is detailed in Section 6.3 with spatiotemporal atlas results presented in Section 6.4. The chapter concludes with a summary in Section 6.5.

6.1 Brain MR Image Database

For this dissertation, MR image data from the CASILab Healthy Subject Database[1] was used. The database has 100 subjects, with twenty subjects in each of five age ranges: 19-29, 30-39, 40-49, 50-59, and 60+ years of age. The subjects are evenly split between females and males. Age distribution of the female and male subjects are shown in Figure 6.1. All subjects were screened for the presence of disease. Additionally, handedness and race were also recorded. For each subject, several 3-Tesla 3D MR images were acquired: a T1-weighted image (FLASH or MPRAGE or both), a T2-weighted image, a magnetic resonance angiogram (MRA), and a diffusion tensor image (DTI). The voxel spacing and dimensions of this data is presented in Figure 6.2. The image acquisition details can be found in Section 3.3.

The multi-modal nature of this database is depicted in Figure 6.3 where the mid-axial slices of the youngest and oldest subjects of both sexes are shown. It is important to note the cross-sectional nature of this data. A spatiotemporal atlas constructed from this database represents many subjects at different ages rather than a single subject at many ages. Therefore, any assessment of volumetric change over time may have the confound of inter-subject variability.

[1]The CASILab Healthy Subject Database was produced by Dr. Elizabeth Bullitt, head of CASILab at UNC Chapel Hill. The work was funded by NIBIB-NIH grant R01 EB000219, 3D Cerebral Vessel Location for Surgical Planning.

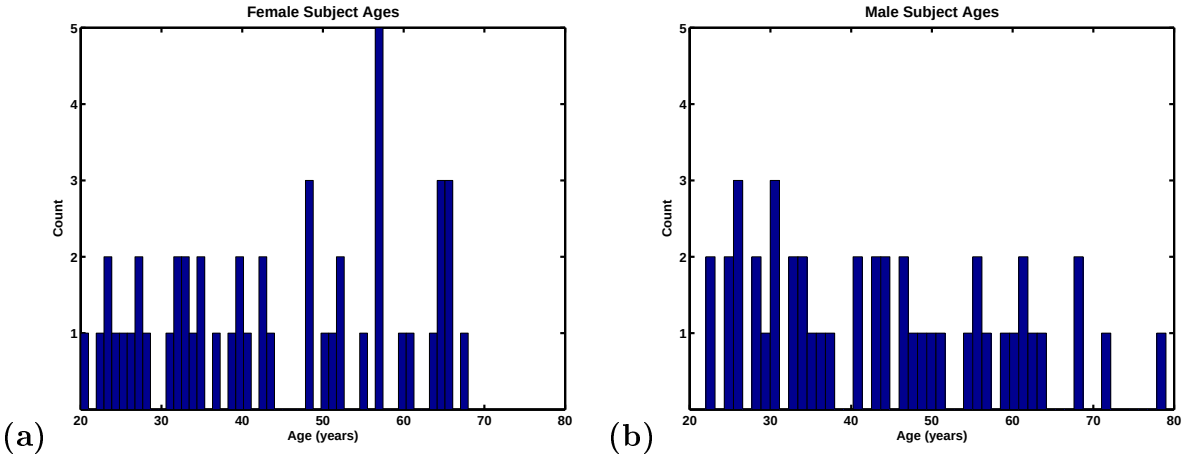


Figure 6.1: Database Subject Ages

Age distribution for the (a) female and (b) male subjects in the database.

	Voxel Spacing (mm)	Dimensions (voxels)
T1-FLASH	$1 \times 1 \times 1$	$176 \times 256 \times 176$
T1-MPRAGE	$1 \times 1 \times 1$	$208 \times 256 \times 128$
T2	$1 \times 1 \times 1$	$192 \times 256 \times 128$
MRA	$0.5 \times 0.5 \times 0.8$	$448 \times 448 \times 128$

Figure 6.2: Image Size

Interior image size (voxel spacing) and exterior image size (dimensions).

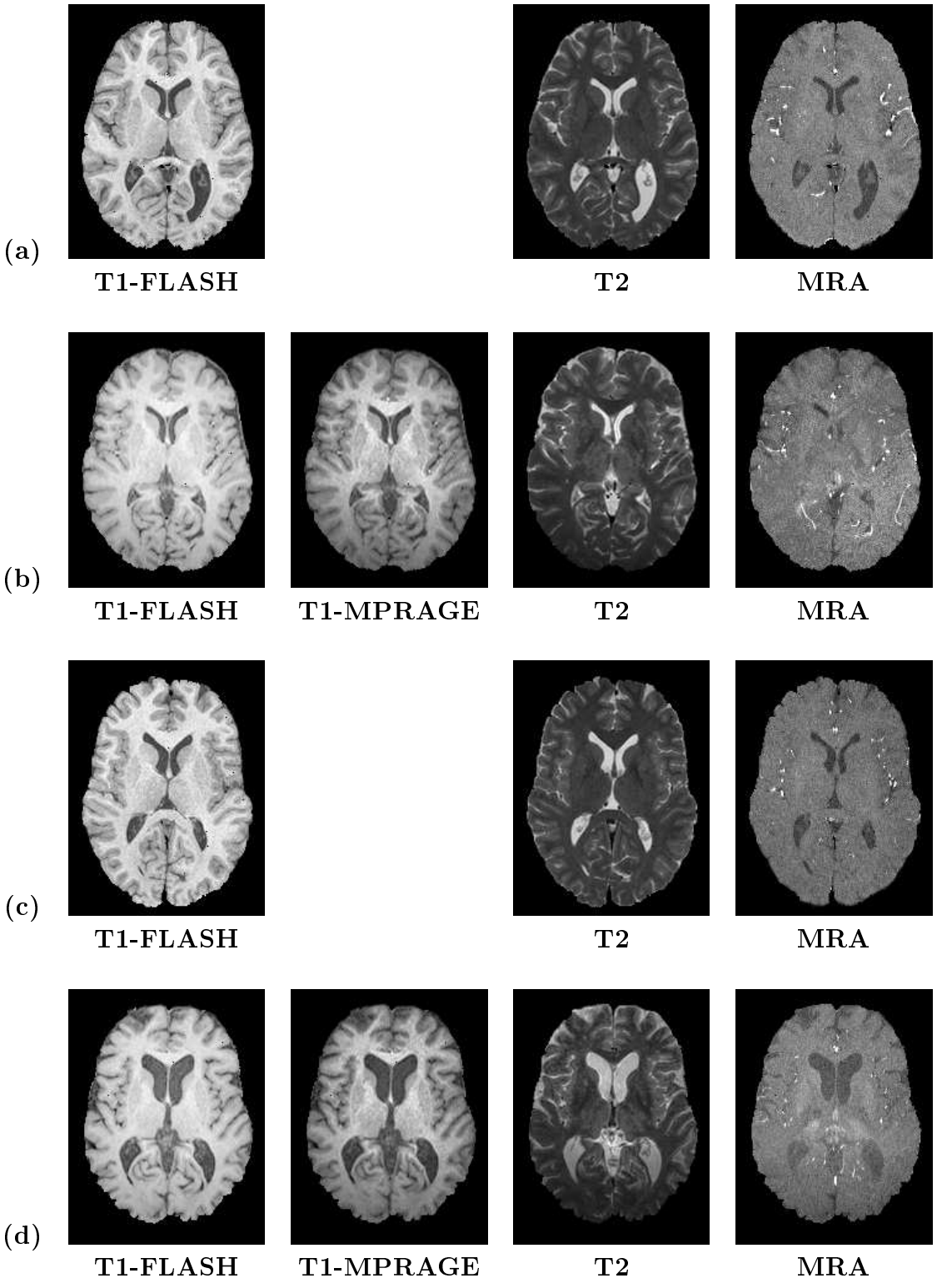


Figure 6.3: Image Database Samples

Mid-axial slices from the (a) youngest (20 year old) subject, (b) the oldest (68 year old) female subject, (c) the youngest (22 year old) male subject, and (d) the oldest (79 year old) male subject.

6.2 Class-Conditional Posterior Maps

To provide direct input to the PMFAtlasBuilder tool, class-conditional posterior maps were created from the T1-weighted and T2-weighted images from the database described in Section 6.1, following a protocol similar to the one presented in a previous study [73] using the same database. Three tissue classes (white matter, gray matter, and cerebrospinal fluid) were defined using an expectation-maximization segmentation method[2] based on an algorithm developed by van Leemput et al. [102, 101]. In addition to the T1 and T2 images, a spatial probabilistic brain atlas [24] representing expert prior knowledge about brain structures was used to drive the segmentation.

Each segmentation was produced in the space of the T1 image resulting in probability maps. For ease of algorithm implementation, an additional background tissue class was added, resulting in four total classes. The class-conditional posterior maps are represented mathematically as

$$p_i(c_j(x)|\bar{I}_i)$$

where the subscript i indicates the subject number, c the tissue class, and $\bar{I}$ the multimodal image set. For the results presented in this chapter, $c = \{$gray matter, white matter, cerebrospinal fluid, and background$\}$, and $\bar{I} = \{$T1 image, T2 image$\}$. Floating point numbers were used to represent the data. Therefore, each posterior map comprises

$$\begin{aligned}\text{number of voxels} \times \text{number of classes} \times \text{data type size} &= 176 \times 256 \times 176 \times 4 \times 4 \text{ Bytes} \\ &= 126,877,696 \text{ Bytes} \\ &= 121 \text{ MBytes.}\end{aligned}$$

6.3 Atlas Formation

This section details the preprocessing of the images in the Healthy Subject Database of Section 6.1 and provides algorithmic analysis of PMFAtlasBuilder. PMFAtlasBuilder was run using the sum-over-class squared error distance described in Section 4.3.1. All preprocessing and experiments described in Section 6.4 were performed in the Neuro-Image Analysis Laboratories (NIAL) at UNC Chapel Hill using a SunFire V40z with

[2]The expection maximization segmention method used to produced the multi-class posteriors was developed by Marcel Prastawa et al. at the University of North Carolina at Chapel Hill [79].

four AMD Opteron CPUs, 16GB of memory, and two 76GB hard drives.

The first step in preprocessing the data involved selecting a coordinate system to which all the class posteriors (produced in Section 6.2) were affinely normalized. For this purpose, a scalar T1-atlas produced from five healthy adult subjects[3] using the unbiased scalar atlas formation method presented in [49] was used. As with the raw T1 and T2 images, this five-subject atlas has isotropic 1mm spacing with dimensions of $160 \times 208 \times 163$ voxels.

Using a program, called areg, for affine registration of multi-modal images, we registered intensity normalized T1 images from each subject to the space of the five-subject T1-atlas. This program was developed by Daniel Rueckert and Julia Schnabel at Imperial College London for their Image Registration Toolkit [86]. This registration tool uses normalized mutual information as the similarity criterion. Cubic spline interpolation was used to resample the resulting images.

The affine transformation parameters from the above registrations were then applied to the class posteriors from Section 6.2, again using cubic spline interpolation. These images were converted from their original *unsigned short* data type format to *single floating point* format for convenience in the algorithm implementation, resulting in class posterior maps of the size

$$\begin{aligned}\text{PMF Data Size} &= \text{Number of Voxels} \times \text{Number of Classes} \times \text{Data Type Size} \\ &= 160 \times 208 \times 163 \times 4 \times 4 \text{ bytes} \\ &= 86,794,240 \text{ bytes} \\ &\approx 82.8 \text{ MBytes.}\end{aligned}$$

6.4 Results

Spatiotemporal atlases were created for both the female and male populations using the PMFAtlasBuilder tool. While preprocessing the database, the segmentation tool failed to produce class-conditional probability maps for one female subject and four male subjects. This appears to be a result of registration failure between the spatial prior used in the segmentation process and the individual subjects. This is most likely due to the affine registration failing to accommodate widely disparate skull shapes.

For each sex, the class-conditional probability maps were ordered by subject age.

[3]These subjects participated in the Schizophrenia First Episode Study at UNC Chapel Hill.

For the female population, this resulted in the set $\{p_i^{female}(c|x)\}_{i=1...49}$ where p_1^{female} and p_{49}^{female} represent the youngest and oldest female subjects respectively. Similarly, for the male population, the ordering resulted in the set $\{p_i^{male}(c|x)\}_{i=1...46}$ where p_1^{male} and p_{46}^{male} represent the youngest and oldest male subjects respectively. The spatiotemporal atlases were then created by building a sequence of atlases, each derived from a fixed number of subjects in that ordering. The number of subjects used to build one of the individual atlases is called the atlas window width w. The jth atlas in the sequence is generated from the sub-population $\{p_j(c|x)\}_{j=i...i+w-1}$.

Due to time constraints atlas stability analysis for this database was not conducted. In the absence of that analysis, spatiotemporal atlases were created using window widths of $w = 10$ and $w = 15$. PMFAtlasBuilder was run for 100 iterations to build each individual atlas. The female and male spatiotemporal atlases for window width $w = 10$ are presented in Figure 6.4 and Figure 6.5 respectively. Similarly, the female and male spatiotemporal atlases for window width $w = 15$ are presented in Figure 6.6 and Figure 6.7 respectively. Greater detail of the youngest and oldest individual female atlases for window widths of $w = 10$ and $w = 15$ are shown in Figures 6.8, 6.9, 6.10, and 6.11. Similarly, greater detail of the youngest and oldest individual male atlases is shown in Figures 6.12, 6.13, 6.14, and 6.15.

The average age of the sub-populations used to create each individual atlas was computed to gain an understanding of how well-spaced in time the 4D spatiotemporal atlases are. These atlas ages and corresponding age differentials are presented in Figure 6.16 and Figure 6.17 for window widths of ten and fifteen subjects respectively.

6.4.1 Spatiotemporal Atlas Stability

To measure spatiotemporal intra atlas stability across constituent individual 3D spatial atlases, the sharpness and population variance of each individual spatial atlas was measured. Using the approach described in Chapter 5, entropy was computed for each class-posterior map using 256 bins. The inter-time entropy trends are shown for female and male spatiotemporal atlases for window widths of $w = 10$ and $w = 15$ in Figures 6.18 and 6.19 respectively. These trends indicate stability with respect to individual atlas sharpness. Individual atlas sub-population variance was measured by considering the distance on the transformations relating each constituent member of the sub-population to the atlas. This distance is the velocity norm in Equation 3.5. These trends for window widths of $w = 10$ and $w = 15$ are presented in Figures 6.20

and 6.21 respectively.

6.4.2 Volumetric Analysis

When viewed as a time sequence of 3D spatial atlases, spatiotemporal atlases provide information regarding volumetric change over time. To analyze this change, per-class volumes were computed for each tissue class using two methods. In the first method, the per-channel volume V_1 of a class c was computed as the sum of class-posterior values over the whole spatial volume Ω,

$$V_1(c) = \sum_{x \in \Omega} p(c|x).$$

This approach assumes partial voluming of tissue classes in each voxel. In the second method, the per-channel volume V_2 of a class c was computed as the sum of maximum a posteriori labels,

$$V_2(c) = \sum_{x \in \Omega} l(c(x))$$

where

$$l(c(x)) = \begin{cases} 1 & \text{where } c = \text{argmin}_{c' \in \{WM, GM, CSF, Background\}} p(c'|x) \\ 0 & \text{otherwise} \end{cases}$$

This approach assumes a voxel is comprised of a single tissue class. Volume trends using both V_1 and V_2 are presented for female and male atlases of width $w = 10$ subjects in Figures 6.22 and 6.23 respectively. Similarly, volume trends for the atlases of width $w = 15$ are shown in Figures 6.24 and 6.25.

The volume trends show a loss of grey matter and an increase in cerebrospinal fluid over time. The global loss of grey matter with age is well documented, for example in aforementioned 465-adult-subject cross-sectional study presented in [30]. In the Good et al. study, the authors report no significant global decrease in white matter volume with age. While this appears to be true for the female data presented in this dissertation it holds for only the younger half of the male data. For the older half of the male population there is a noticeable decrease in white matter volume. There also appears to be a general decrease in the total intracranial volume with age for the entire

male population. Additionally, Good et al. report a steeper decline in grey matter volume in males than in females whereas, in this dissertation, the grey matter loss appears to be similar for the female and male populations. One possible explanation for the greater than expected white matter loss and lower than expected grey matter loss for second half of the male population is the progressively poorer contrast over time between grey matter and white matter. For example, consider occipital region in the T1-FLASH and T2 images for the oldest male in Figure 6.3.

There is agreement between the white matter volume of the individual atlases and the average white matter volume of the sub-populations from which those atlases where created. While this consistency holds for the gray matter class, it does not hold for the cerebrospinal fluid class. The computed cerebrospinal fluid volumes for the individual atlases fall well below their corresponding sub-population averages. This is due to the thinning and, in places, destruction of cortical cerebrospinal fluid voxels during the linear interpolation used to resample the images during the registration process. Since much of the cerebrospinal fluid is cortical, this effect is quite pronounced. The effect of linear interpolation on atlas volumes can be studied by computing the volume of a single class-posterior map at scale. The class-posterior map representing the youngest female subject was blurred with a Gaussian kernel with $\sigma = 0.5$ and $\sigma = 1.0$ voxels. The class volumes of these two blurred class-posterior maps are compared with the class volumes of the original class-posterior maps in Figure 6.26. As the class-posterior maps are blurred, the cerebrospinal fluid volume decreases.

Local volumetric change can be studied by analyzing the logarithm of Jacobian maps derived from the transformation relating the youngest and oldest spatial atlases within a given spatiotemporal atlas. Figures 6.27 and 6.28 show these log-Jacobian maps for the female and male populations for window widths $w = 10$ and $w = 15$ respectively. In these images, blue represents a volumetric contraction, green no change, and red volumetric expansion.

6.5 Summary

In this chapter, the multi-class posterior atlas formation method was applied to a database of multi-modal images from ninety-five adult brains as part of a healthy aging study. In this study, 4D spatiotemporal atlases were created for the male and female populations. This work is unique in that Fréchet mean atlases were created for a number of time points. These mean atlases suggest an approximation for the

line of best fit class posteriors over time. Based on the results in Chapter 5, sliding windows of ten and fifteen subjects were used to ensure temporal smoothness. All of this was facilitated by the good age distribution of the subjects in the database used. Volumetric analysis of white-matter, grey-matter, and cerebrospinal fluid change over time were consistent with results from previous studies involving large databases. The use of sharp spatiotemporal atlases provides an opportunity to analyze local volumetric change over time.

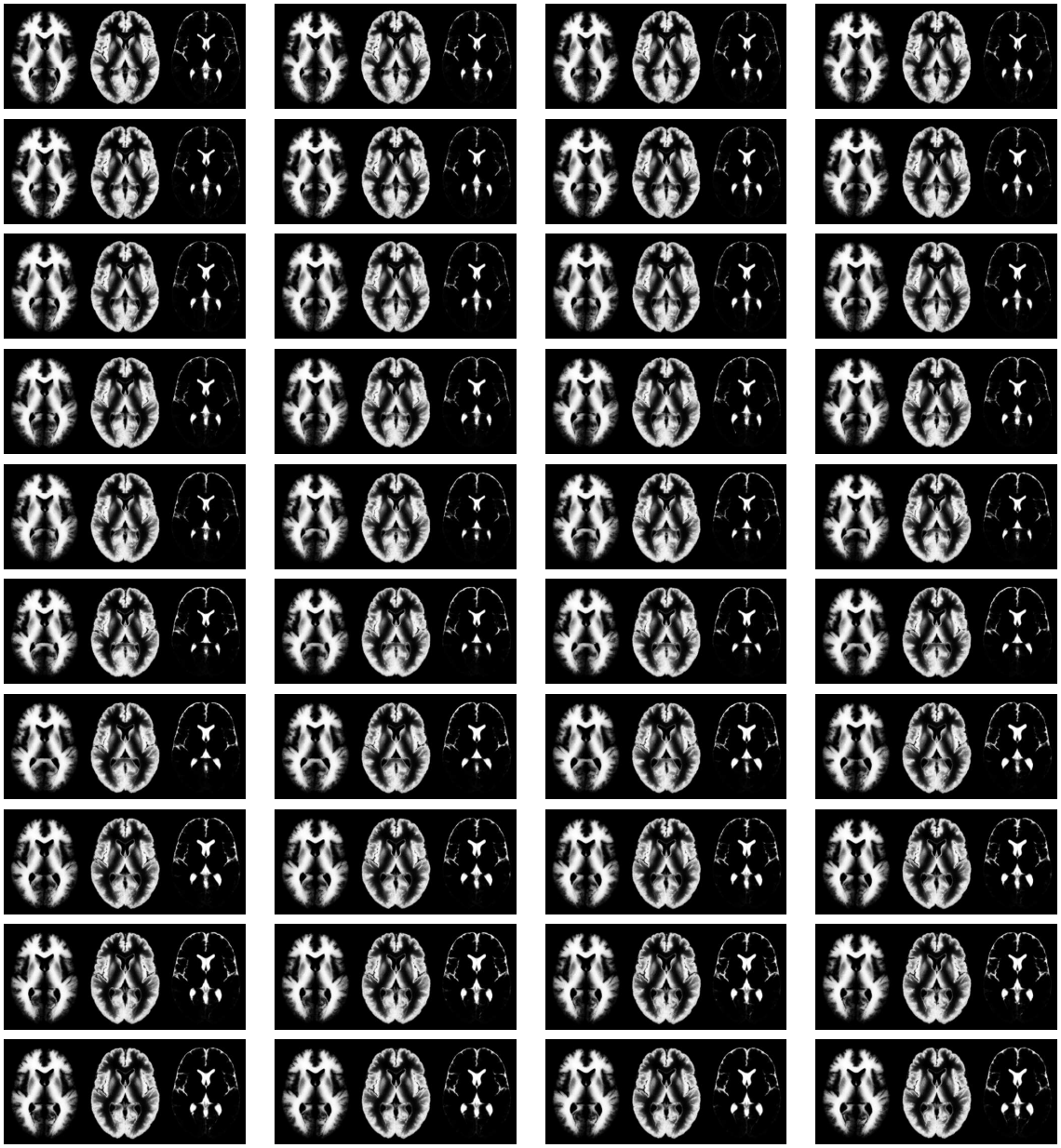


Figure 6.4: Female Spatiotemporal Atlas, $w = 10$

Mid-axial slice view of the forty individual female atlases generated using a window width of ten subjects.

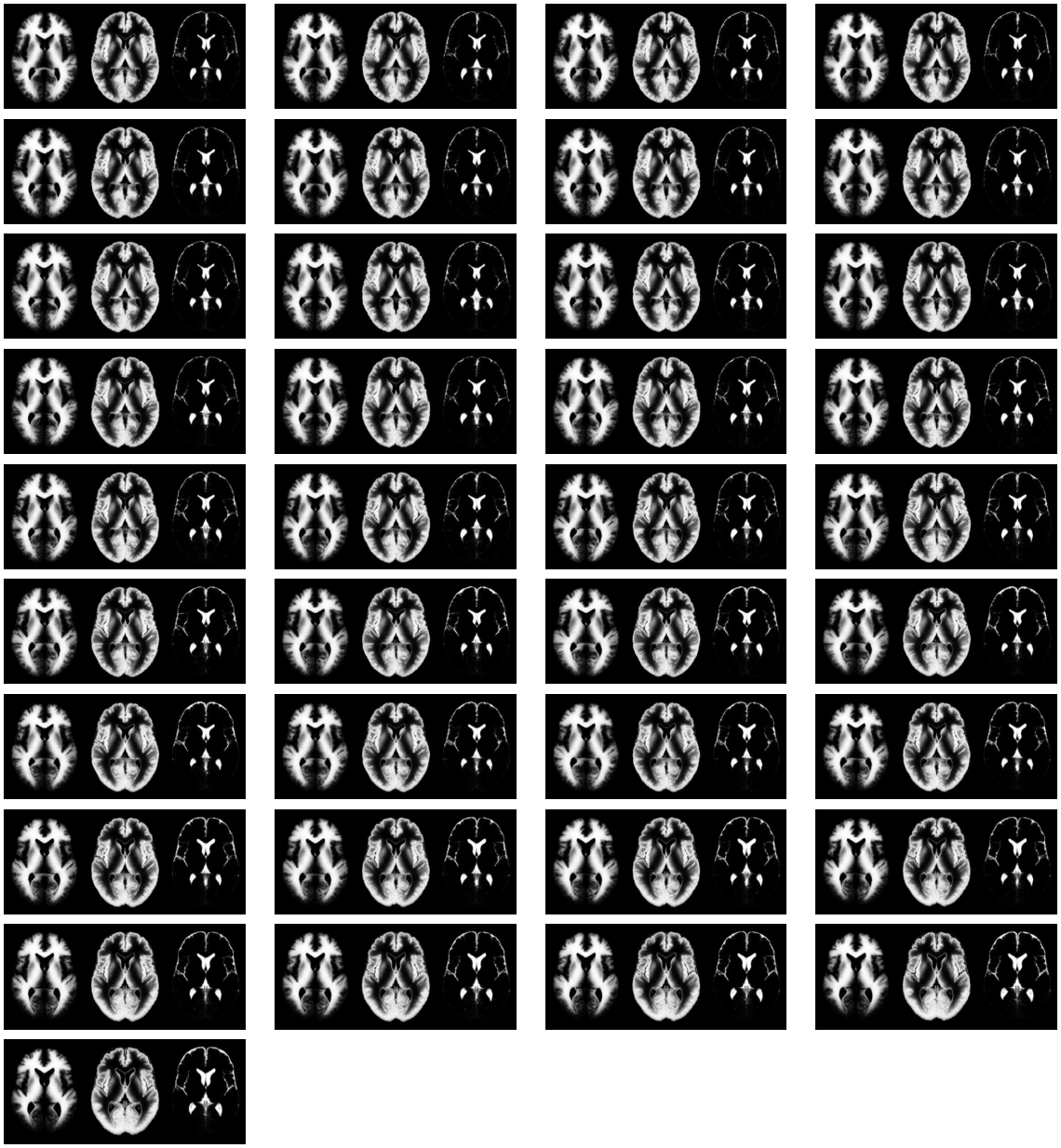


Figure 6.5: Male Spatiotemporal Atlas, $w = 10$

Mid-axial slice view of the thirty-seven individual male atlases generated using a window width of ten subjects.

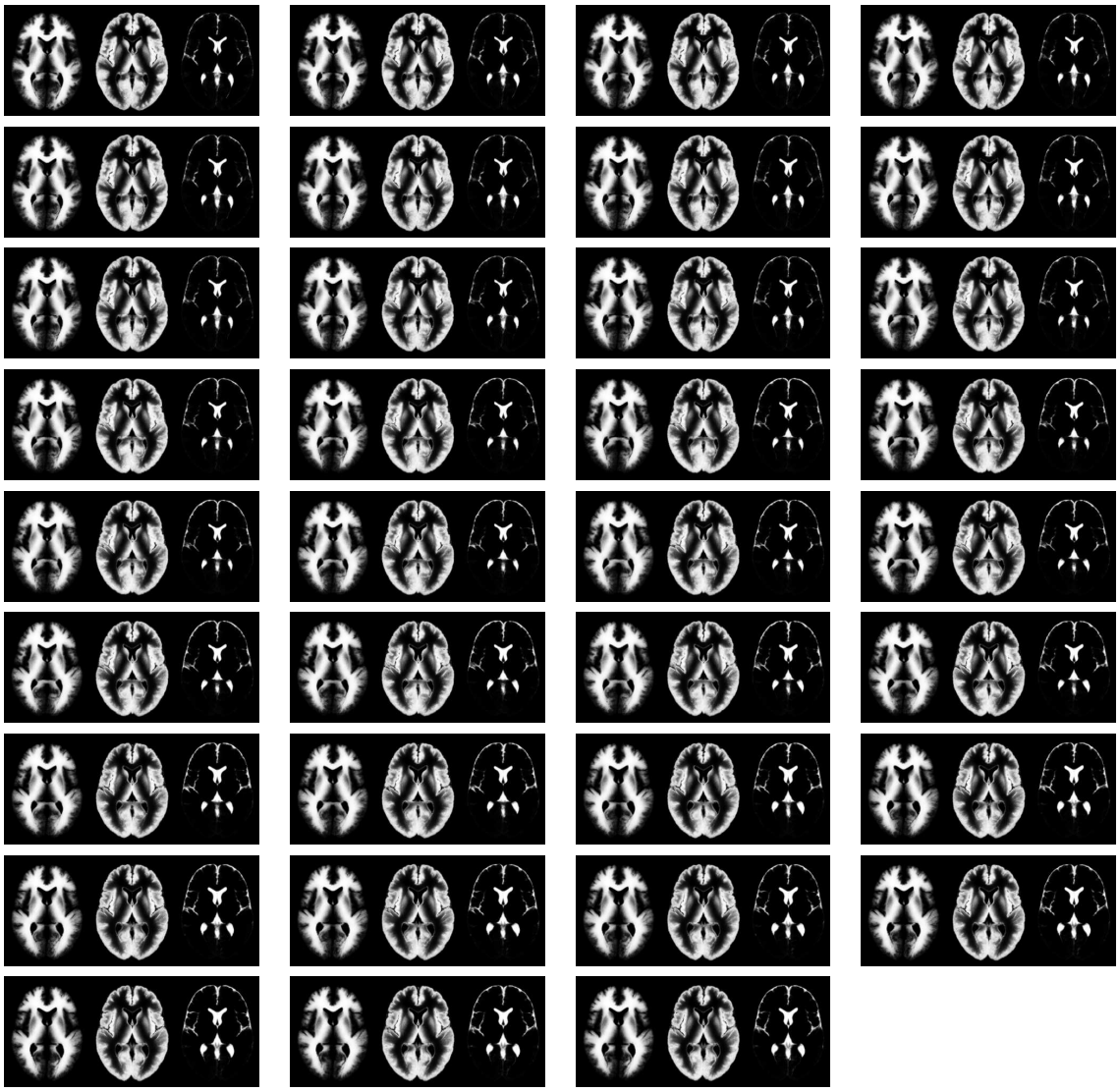


Figure 6.6: Female Spatiotemporal Atlas, $w = 15$

Mid-axial slice view of the thirty-five individual female atlases generated using a window width of fifteen subjects.

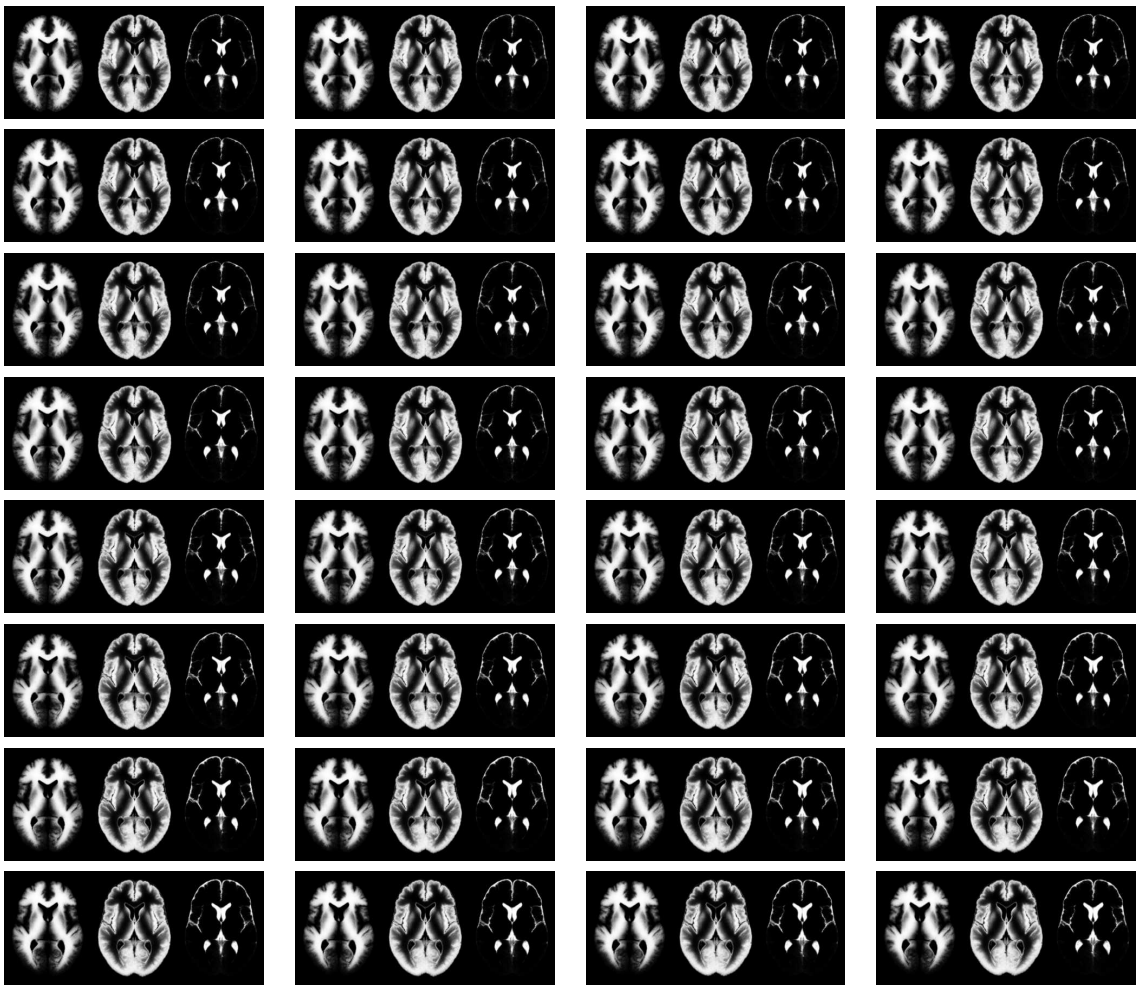
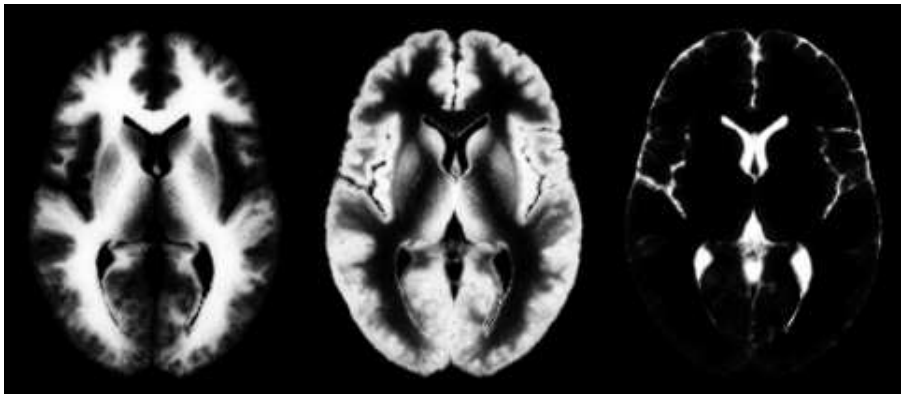
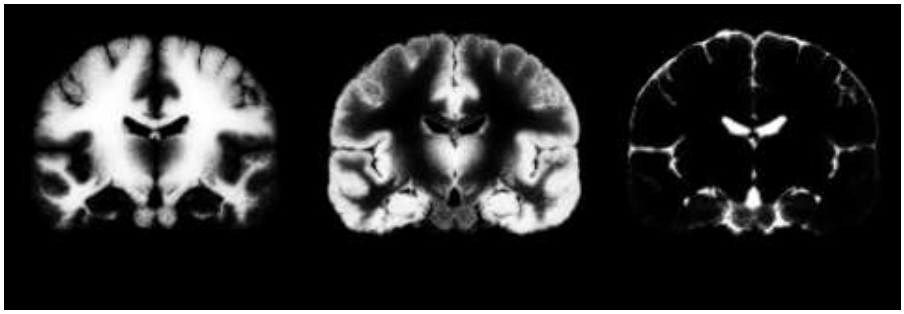


Figure 6.7: Male Spatiotemporal Atlas, $w = 15$

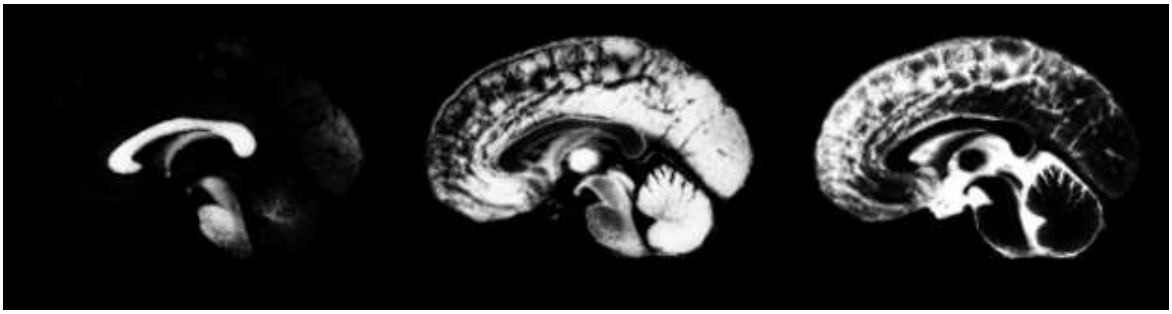
Mid-axial slice view of the thirty-two individual male atlases generated using a window width of fifteen subjects.



(a) Mid-Axial Slice



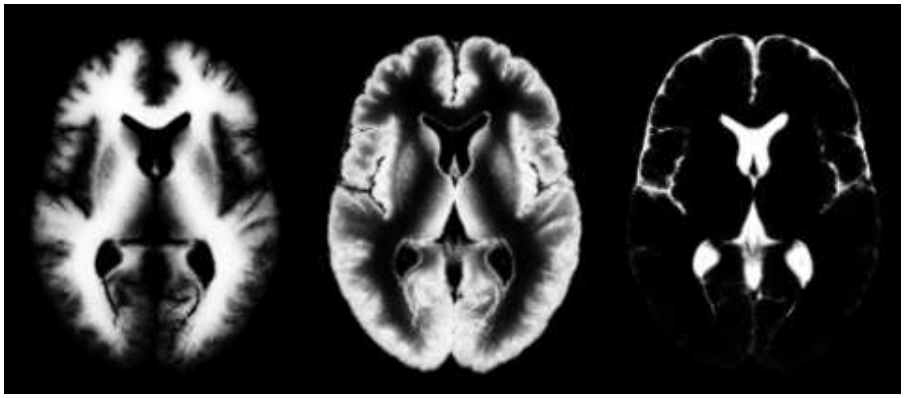
(b) Mid-Coronal Slice



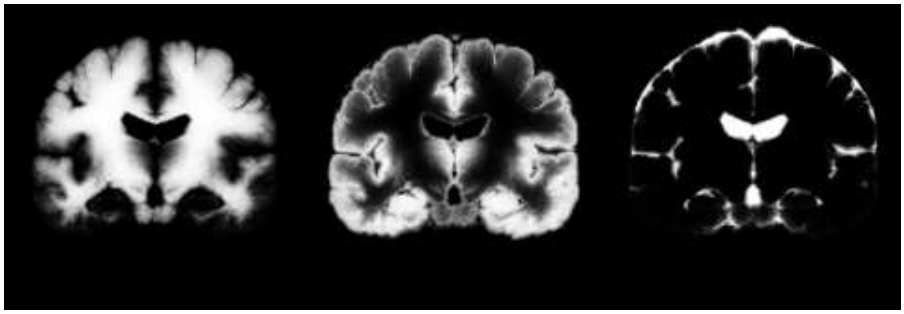
(c) Mid-Sagittal Slice

Figure 6.8: Youngest Female Atlas, $w = 10$

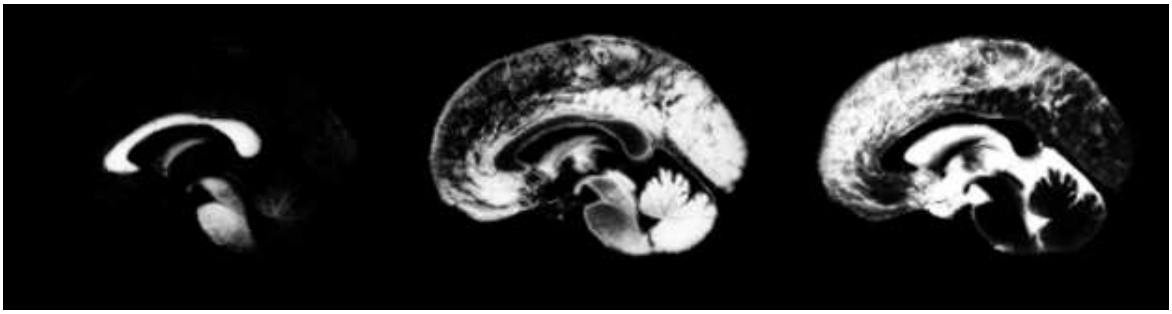
The mid-axial (a), mid-coronal (b), and mid-sagittal (c) slices of the atlas built from the class posteriors representing the ten youngest females in the database.



(a) Mid-Axial Slice



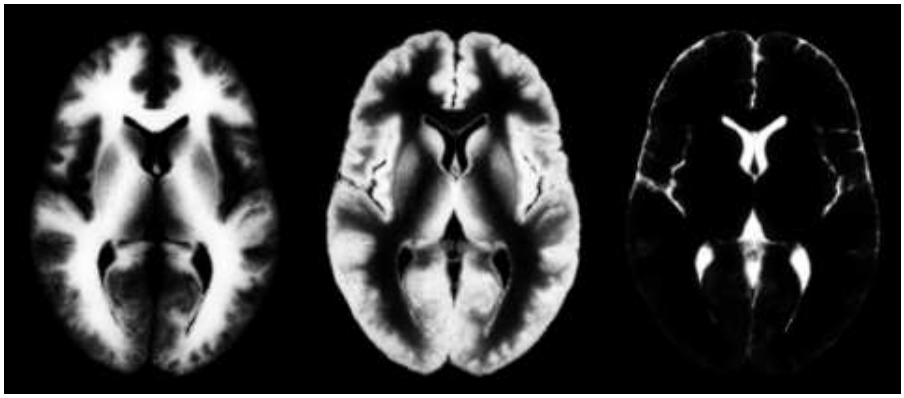
(b) Mid-Coronal Slice



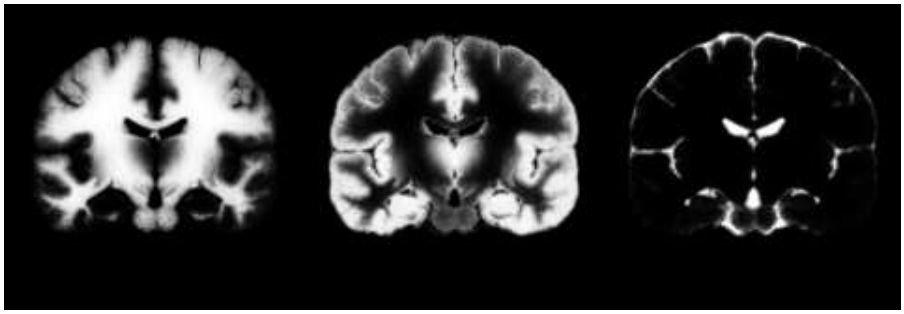
(c) Mid-Sagittal Slice

Figure 6.9: Oldest Female Atlas, $w = 10$

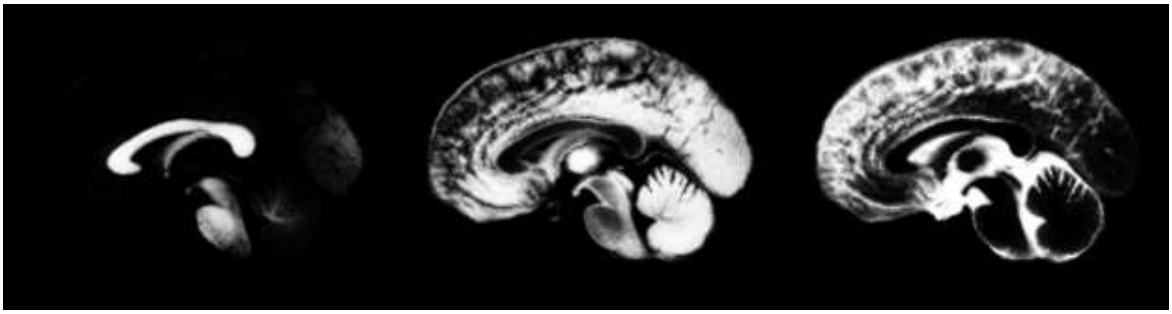
The mid-axial (a), mid-coronal (b), and mid-sagittal (c) slices of the atlas built from the class posteriors representing the ten oldest females in the database.



(a) Mid-Axial Slice



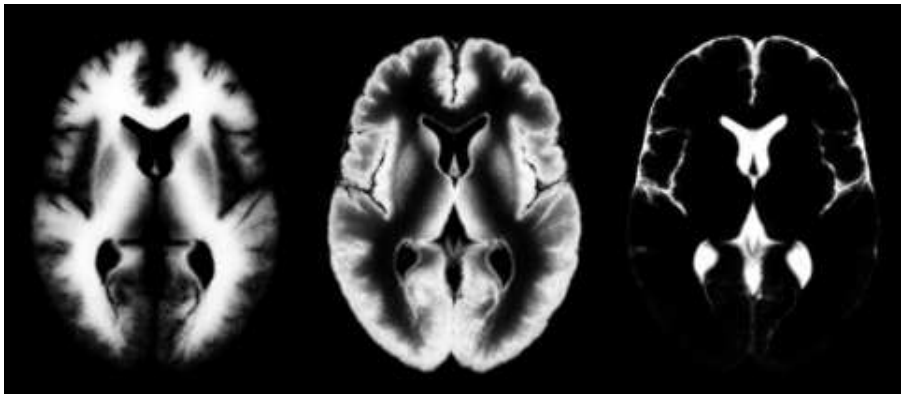
(b) Mid-Coronal Slice



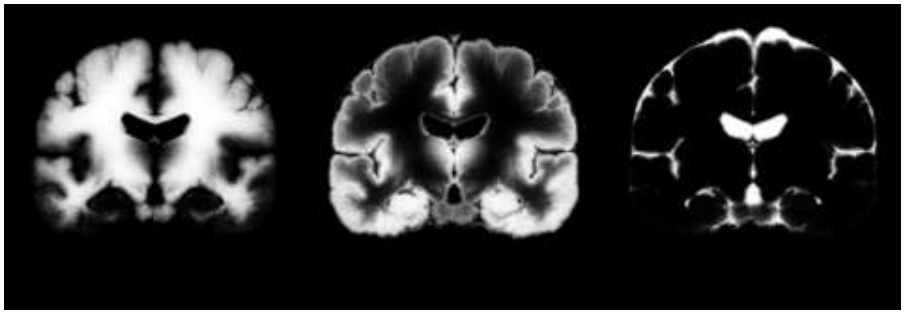
(c) Mid-Sagittal Slice

Figure 6.10: Youngest Female Atlas, $w = 15$

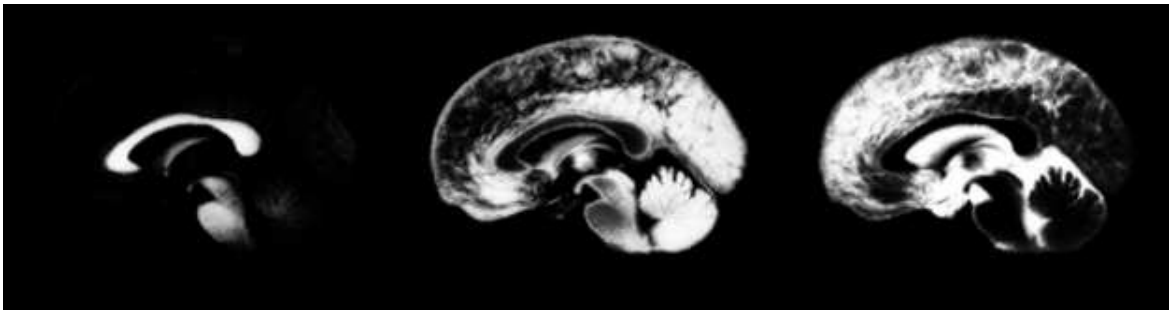
The mid-axial (a), mid-coronal (b), and mid-sagittal (c) slices of the atlas built from the class posteriors representing the ten youngest females in the database.



(a) Mid-Axial Slice



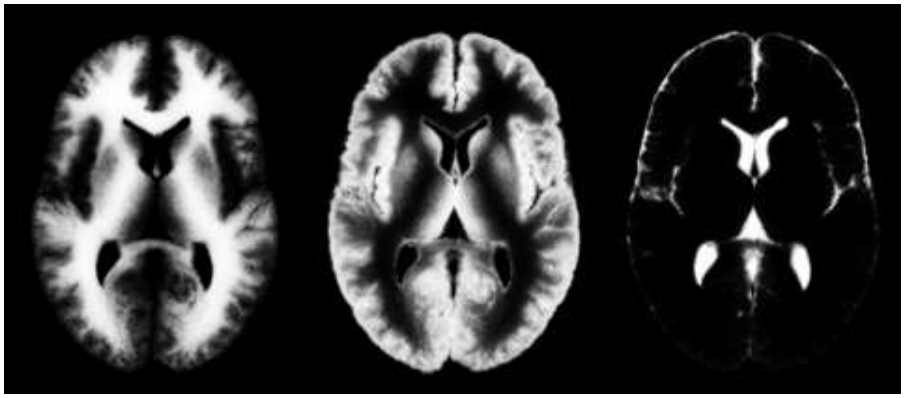
(b) Mid-Coronal Slice



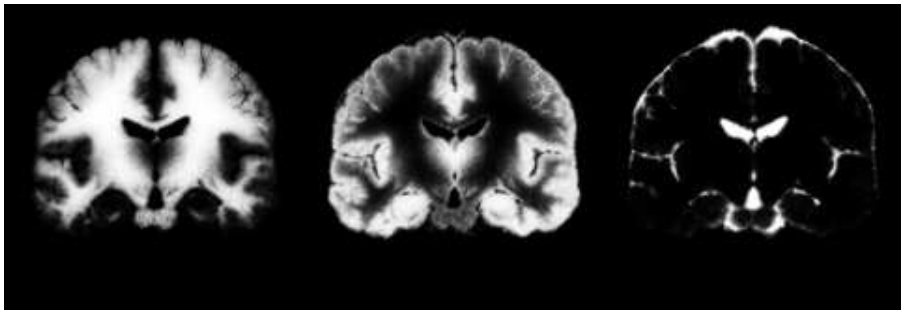
(c) Mid-Sagittal Slice

Figure 6.11: Oldest Female Atlas, $w = 15$

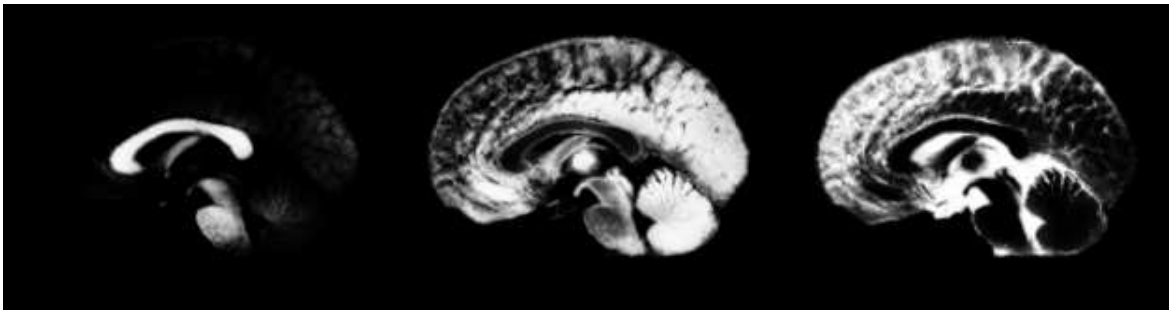
The mid-axial (a), mid-coronal (b), and mid-sagittal (c) slices of the atlas built from the class posteriors representing the ten oldest females in the database.



(a) Mid-Axial Slice



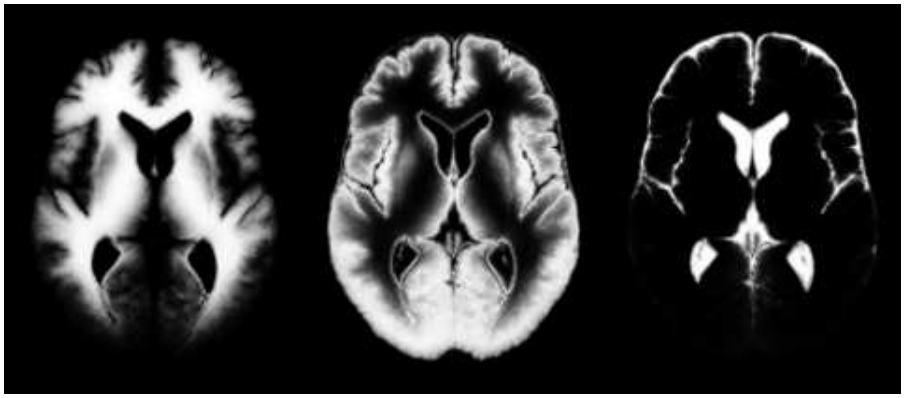
(b) Mid-Coronal Slice



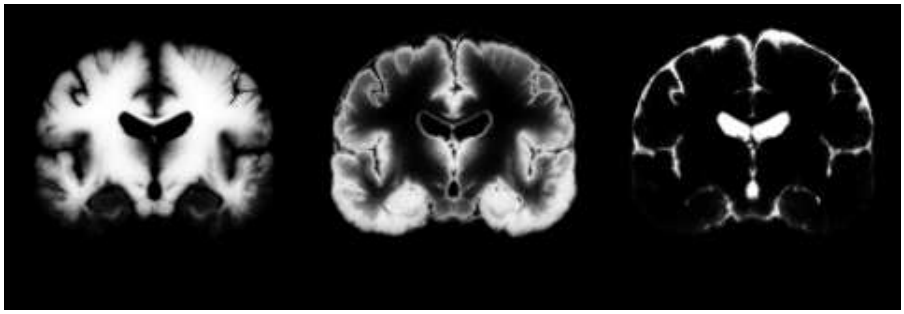
(c) Mid-Sagittal Slice

Figure 6.12: Youngest Male Atlas, $w = 10$

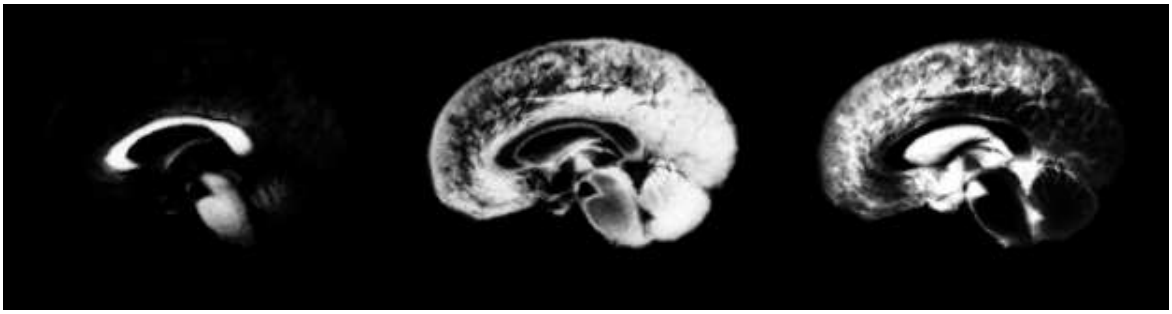
The mid-axial (a), mid-coronal (b), and mid-sagittal (c) slices of the atlas built from the class posteriors representing the ten youngest males in the database.



(a) Mid-Axial Slice



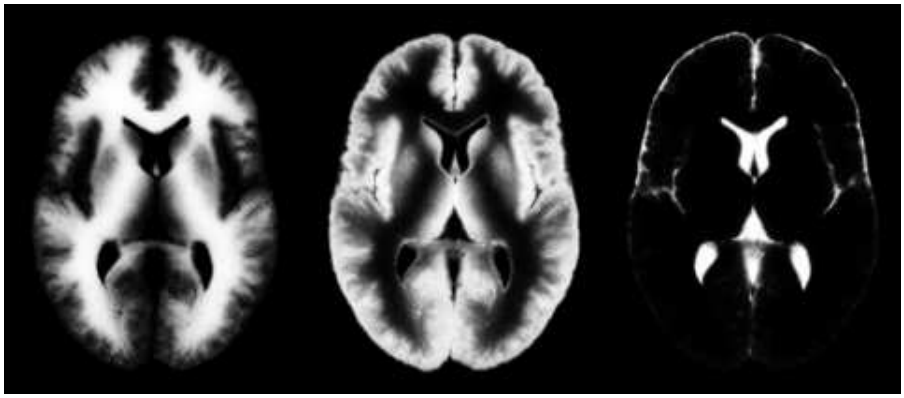
(b) Mid-Coronal Slice



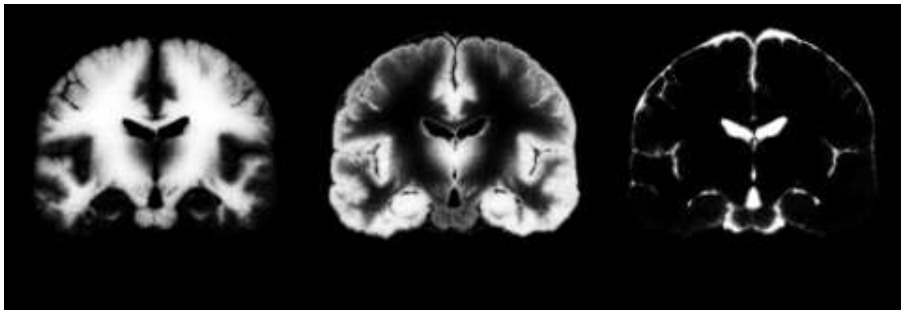
(c) Mid-Sagittal Slice

Figure 6.13: Oldest Male Atlas, $w = 10$

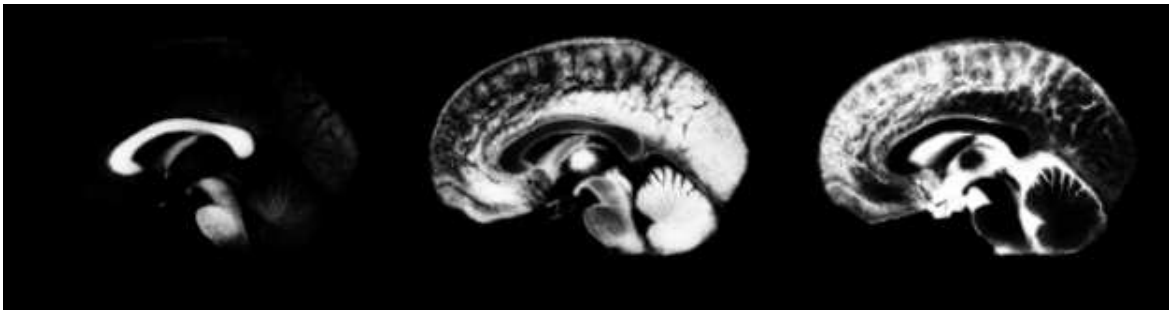
The mid-axial (a), mid-coronal (b), and mid-sagittal (c) slices of the atlas built from the class posteriors representing the ten oldest males in the database.



(a) Mid-Axial Slice



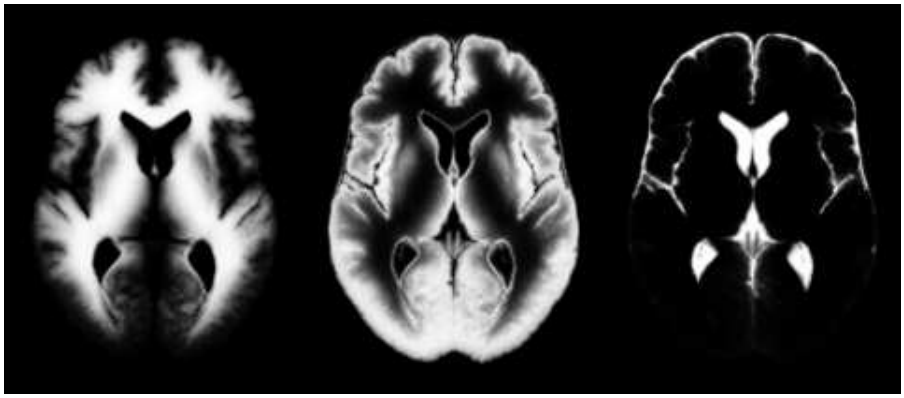
(b) Mid-Coronal Slice



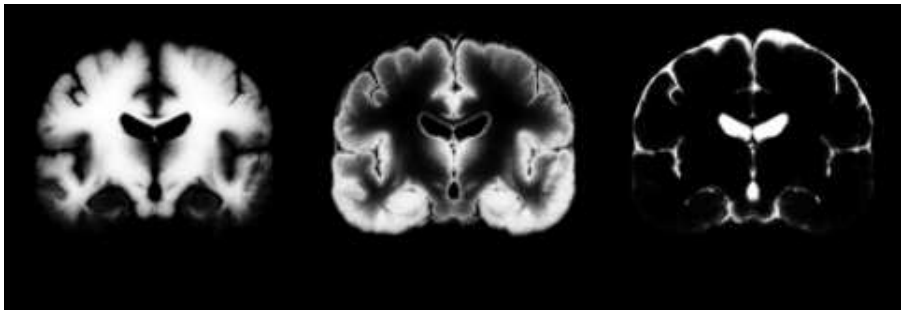
(c) Mid-Sagittal Slice

Figure 6.14: Youngest Male Atlas, $w = 15$

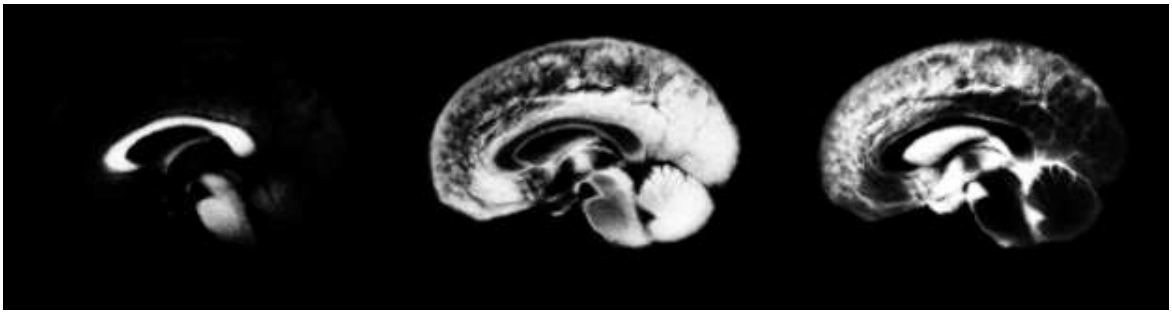
The mid-axial (a), mid-coronal (b), and mid-sagittal (c) slices of the atlas built from the class posteriors representing the ten youngest males in the database.



(a) Mid-Axial Slice



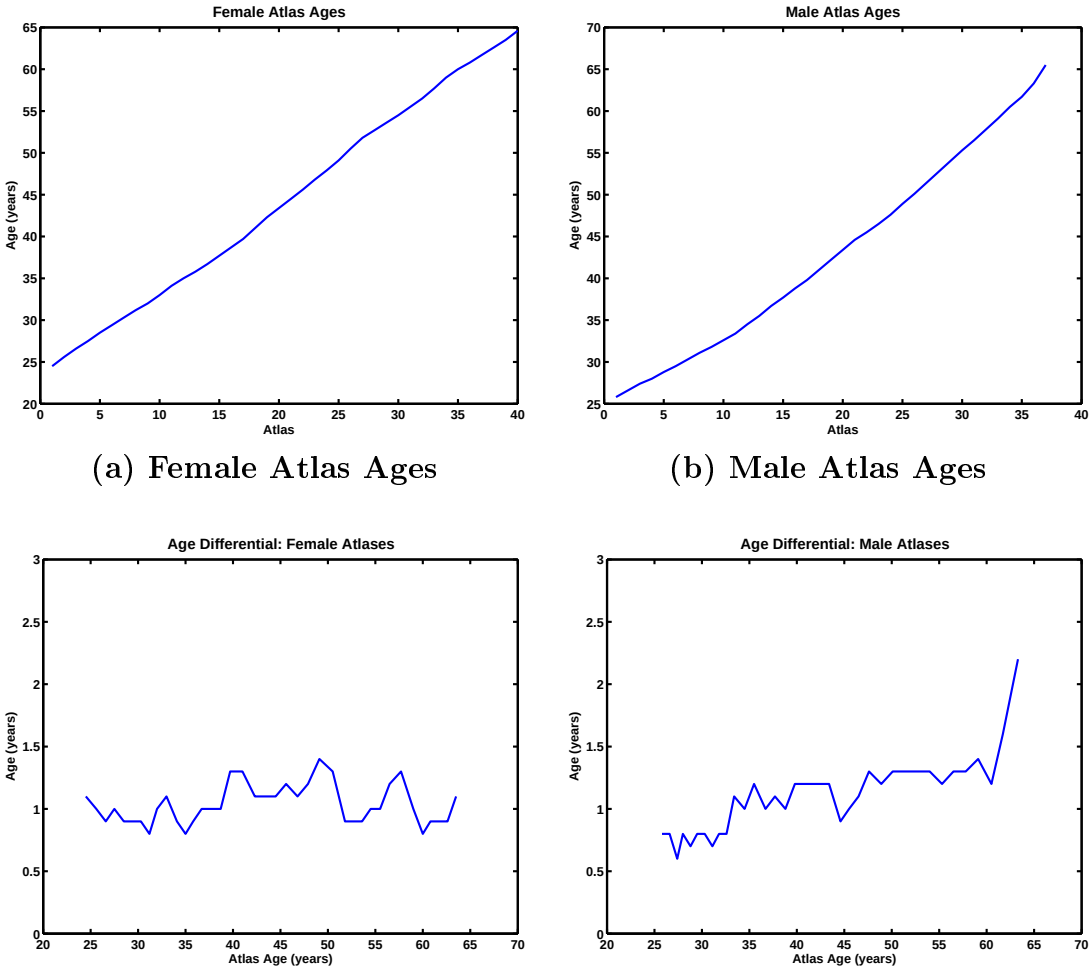
(b) Mid-Coronal Slice



(c) Mid-Sagittal Slice

Figure 6.15: Oldest Male Atlas, $w = 15$

The mid-axial (a), mid-coronal (b), and mid-sagittal (c) slices of the atlas built from the class posteriors representing the ten oldest males in the database.

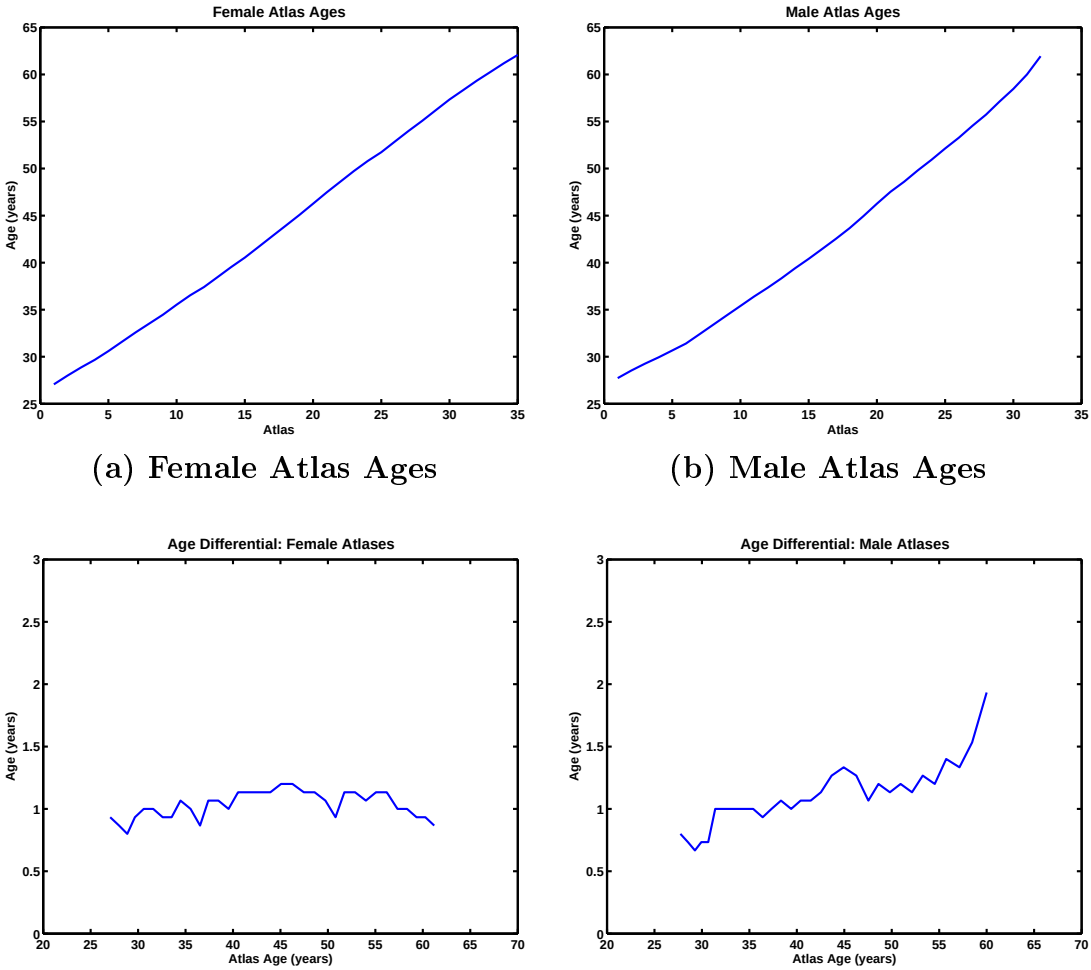


(a) Female Atlas Ages **(b) Male Atlas Ages**

(c) Female Atlas Age Differentials **(d) Male Atlas Age Differentials**

Figure 6.16: Atlas Age Trends, $w = 10$

Average age of subjects used to create the (a) female atlases and the (b) male atlases. Differential age for plots (a) and (b) is shown in (c) and (d) respectively.



(a) Female Atlas Ages **(b) Male Atlas Ages**

(c) Female Atlas Age Differentials **(d) Male Atlas Age Differentials**

Figure 6.17: Atlas Age Trends, $w = 15$

Average age of subjects used to create the (a) female atlases and the (b) male atlases. Differential age for plots (a) and (b) is shown in (c) and (d) respectively.

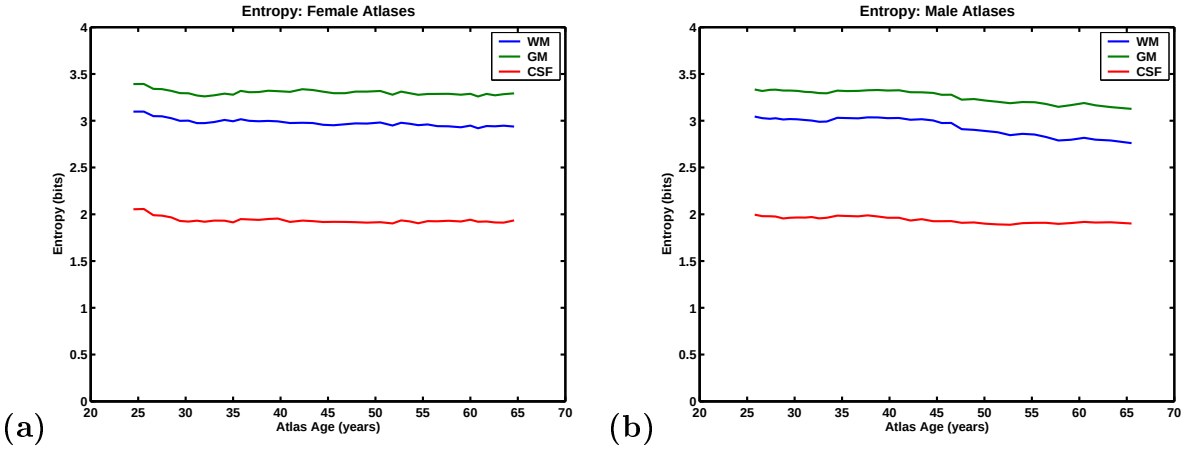


Figure 6.18: Entropy Trends, $w = 10$

Class-posterior entropy trends for the female (a) and male (b) atlases.

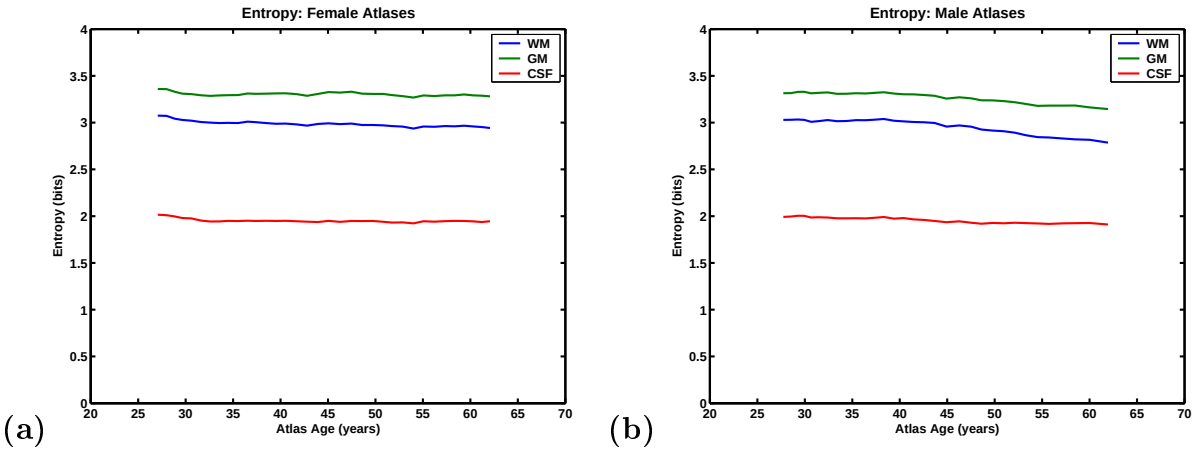


Figure 6.19: Entropy Trends, $w = 15$

Class-posterior entropy trends for the female (a) and male (b) atlases.

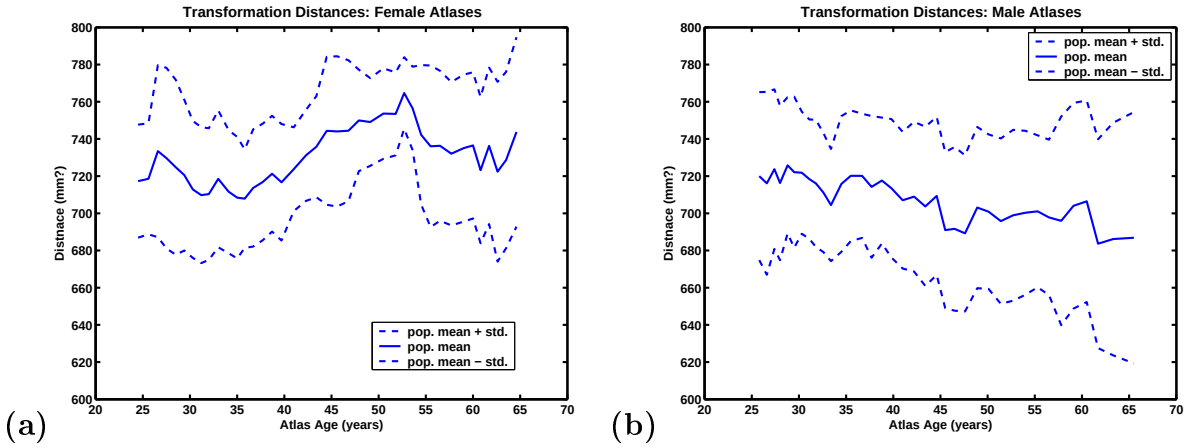


Figure 6.20: Velocity Norm Trends, $w = 10$

Velocity norm trends for the female (a) and male (b) atlases.

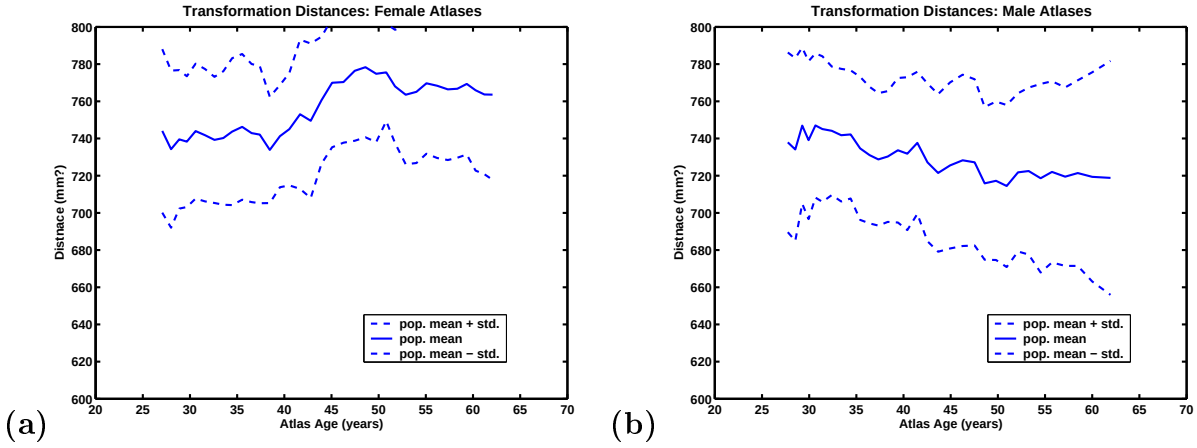


Figure 6.21: Velocity Norm Trends, $w = 15$

Velocity norm trends for the female (a) and male (b) atlases.

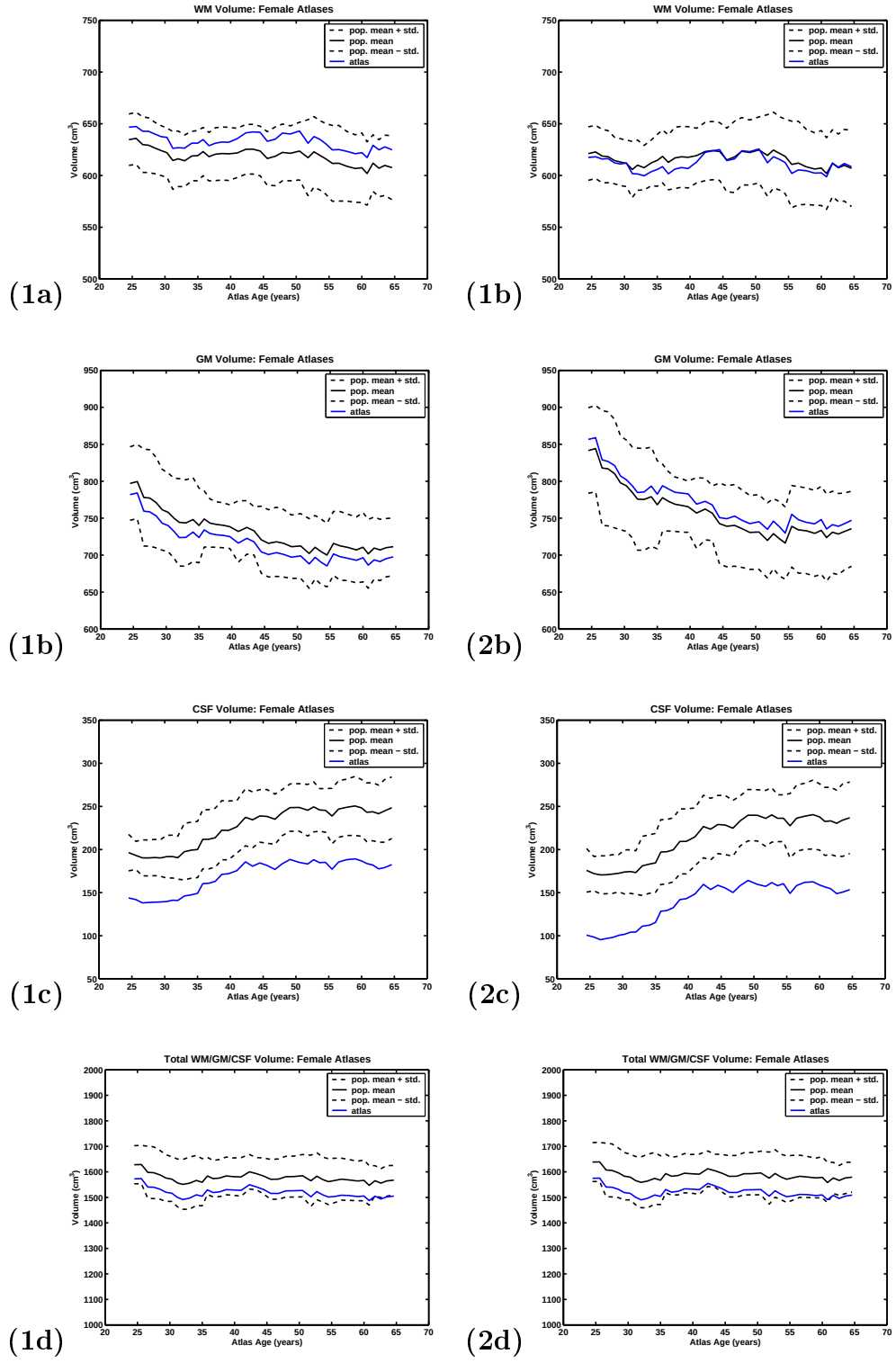


Figure 6.22: Female Volume Trends, $w = 10$

Volume trends for (1a-1d) V_1 and (2a-2d) V_2 for the female atlases of width of ten subjects for (1a&2a) white matter, (1b&2b) grey matter, (1c&2c) cerebrospinal fluid, and (1d&2d) total brain volume.

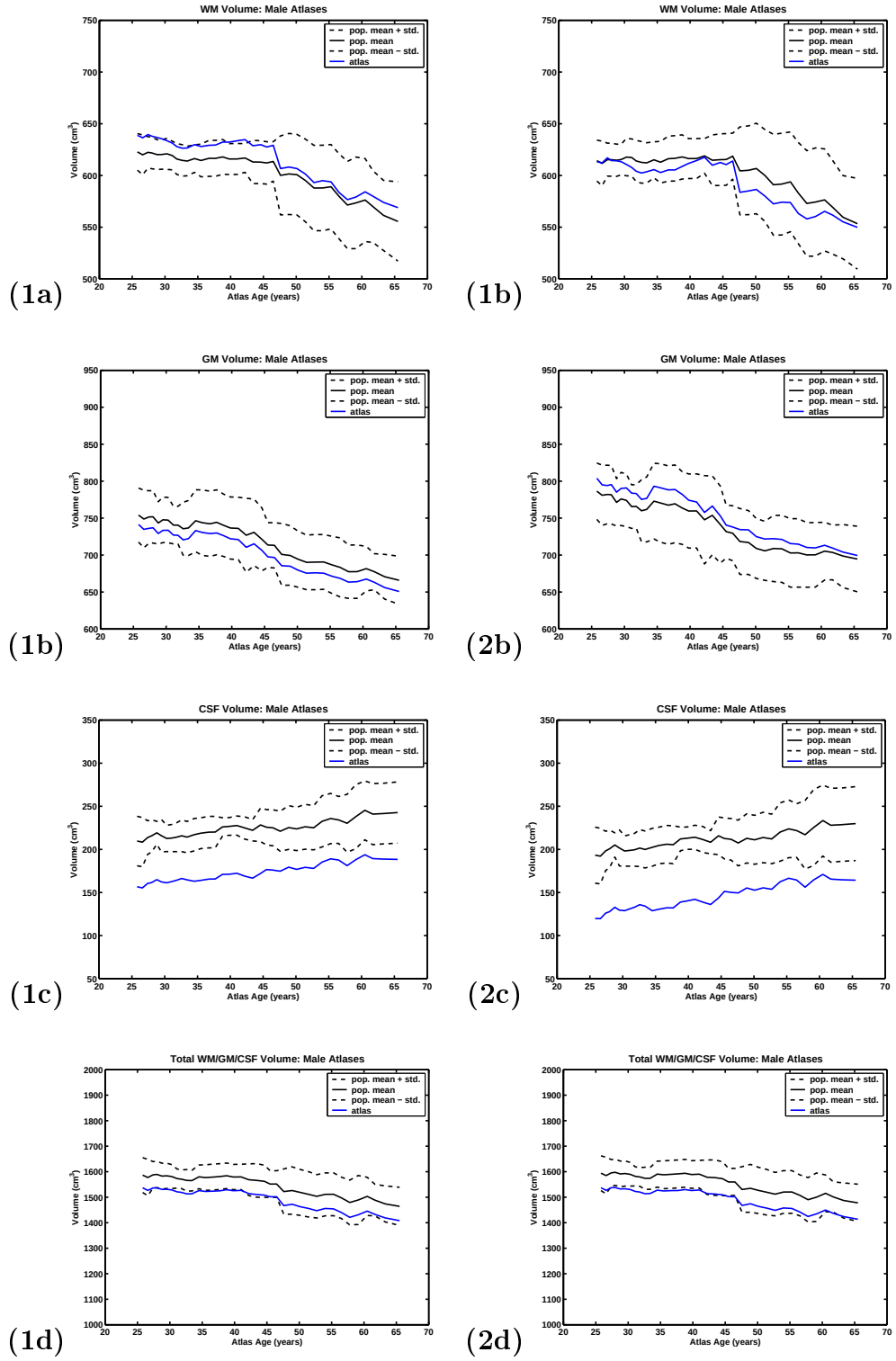


Figure 6.23: Male Volume Trends, $w = 10$

Volume trends for (1a-1d) V_1 and (2a-2d) V_2 for the male atlases of width of ten subjects for (1a&2a) white matter, (1b&2b) grey matter, (1c&2c) cerebrospinal fluid, and (1d&2d) total brain volume.

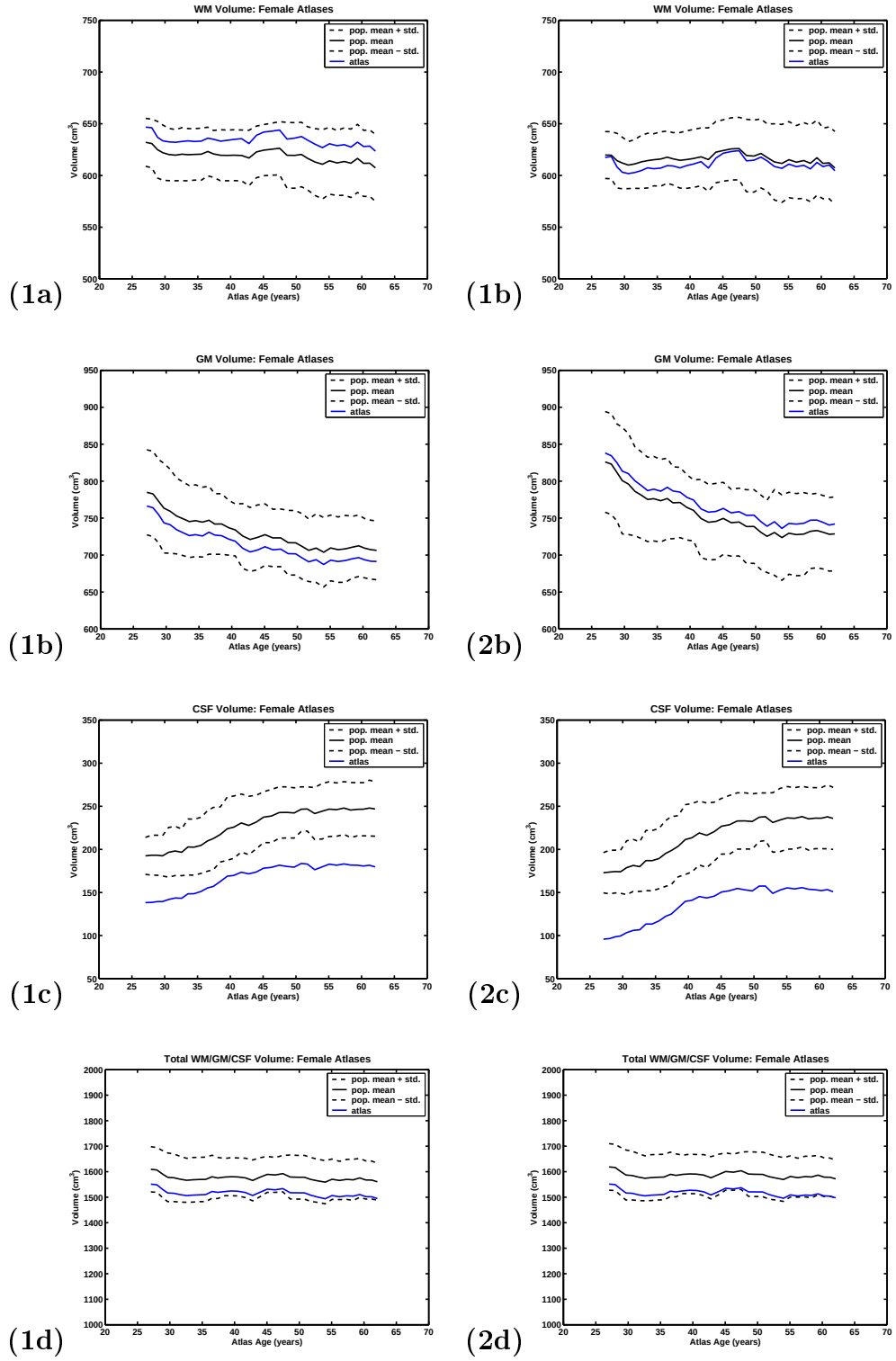


Figure 6.24: Female Volume Trends, $w = 15$

Volume trends for (1a-1d) V_1 and (2a-2d) V_2 for the female atlases of width of fifteen subjects for (1a&2a) white matter, (1b&2b) grey matter, (1c&2c) cerebrospinal fluid, and (1d&2d) total brain volume.

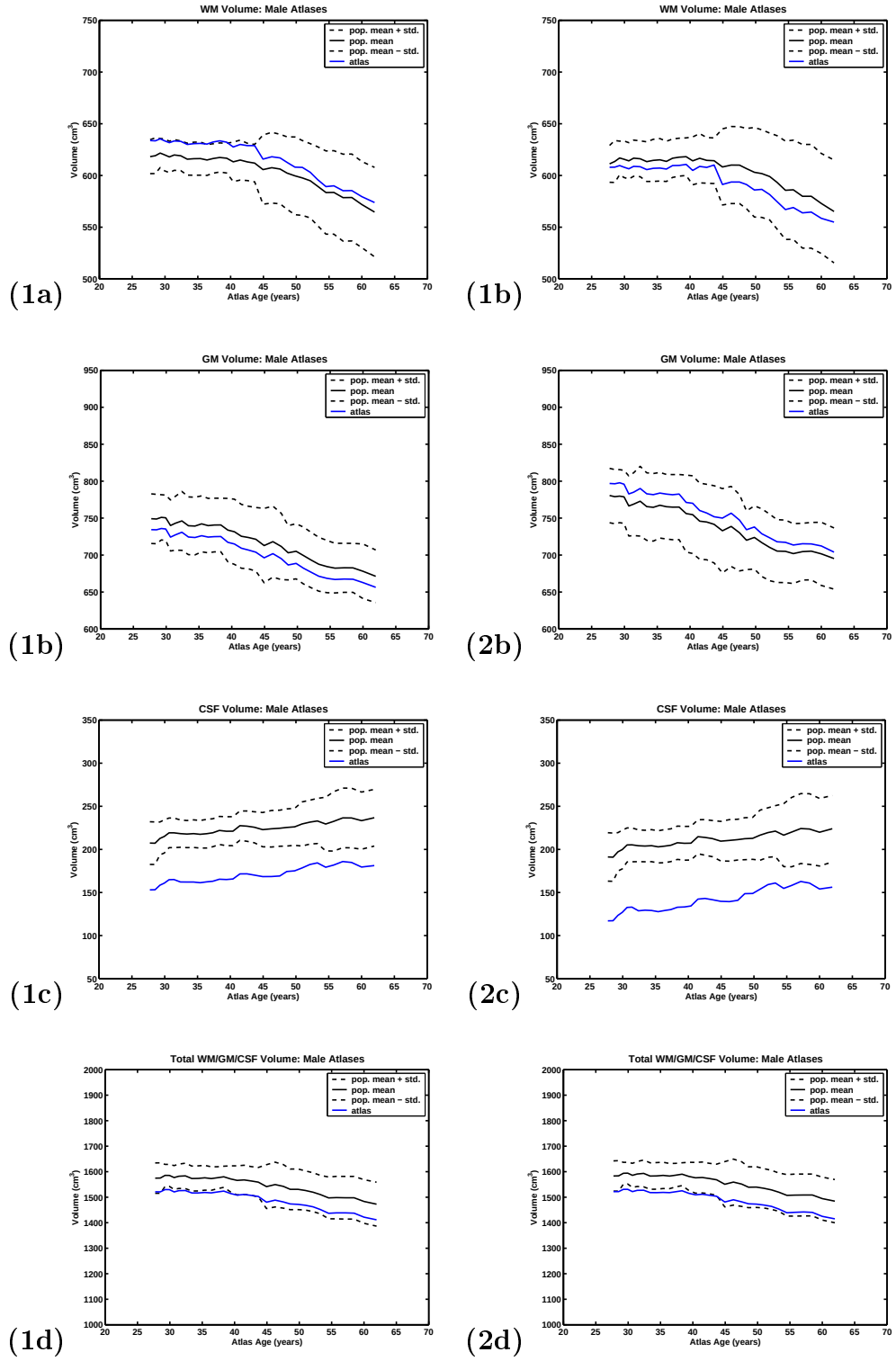


Figure 6.25: Male Volume Trends, $w = 15$

Volume trends for (1a-1d) V_1 and (2a-2d) V_2 for the female atlases of width of fifteen subjects for (1a&2a) white matter, (1b&2b) grey matter, (1c&2c) cerebrospinal fluid, and (1d&2d) total brain volume.

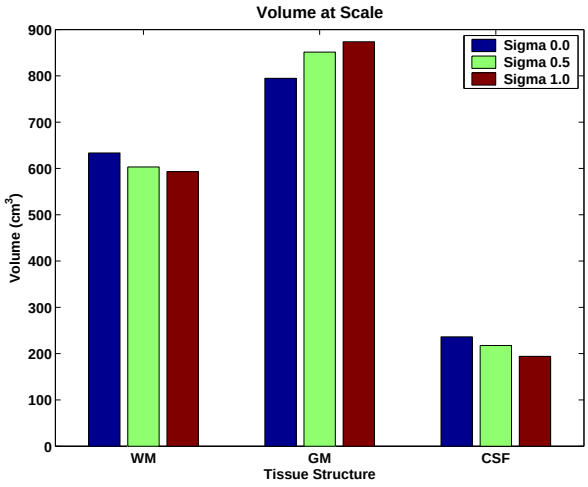


Figure 6.26: Volume at Scale

Class volumes for the youngest female and two blurred versions of the same.

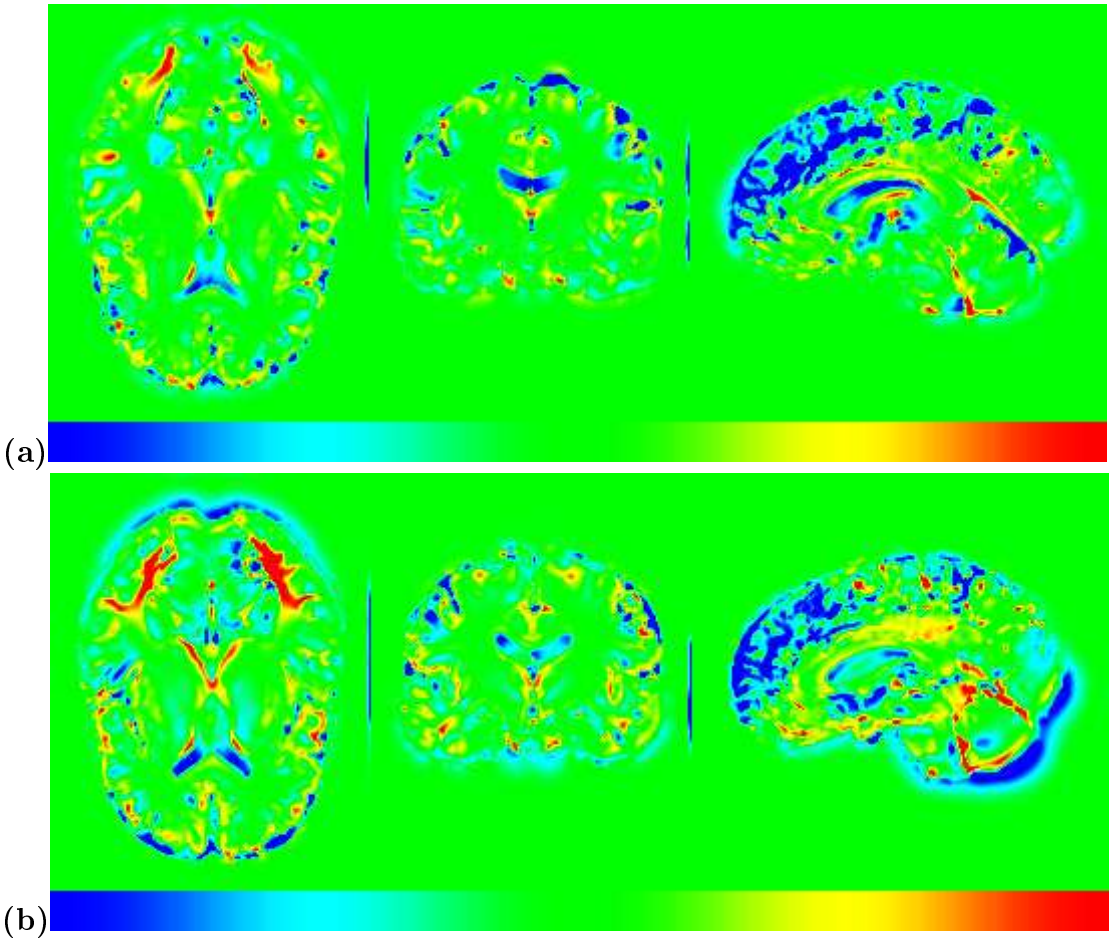


Figure 6.27: Log-Jacobian Map: Oldest to Youngest, $w = 10$

Log-Jacobian maps of the transformation relating the coordinate spaces of the (a) oldest to youngest female and (b) oldest to youngest male. The color in these figures represents volumetric change, with respect to the original, of 32% or less for blue, 100% for green, 316% or more for red.

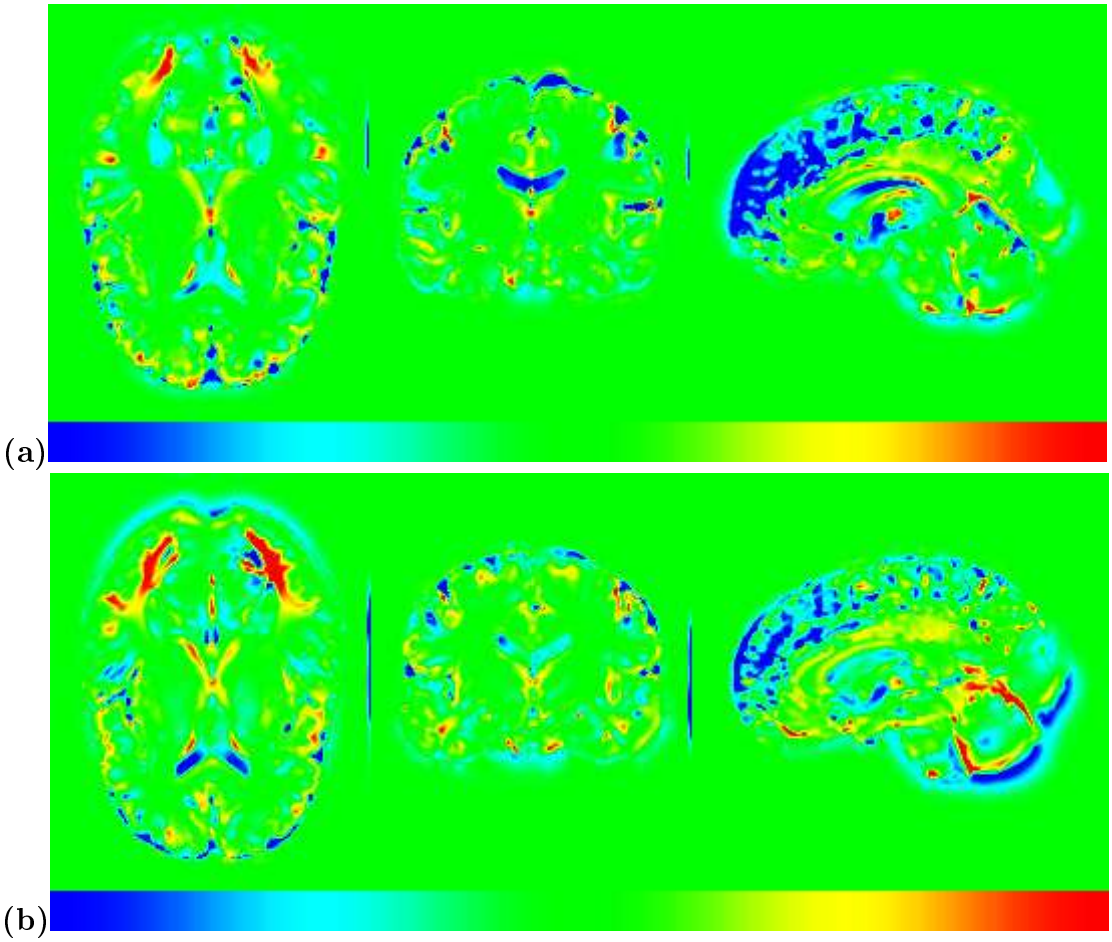


Figure 6.28: Log Jacobian Map: Oldest to Youngest, $w = 15$

Log-Jacobian maps of the transformation relating the coordinate spaces of the (a) oldest to youngest female and (b) oldest to youngest male. The color in these figures represents volumetric change, with respect to the original, of 32% or less for blue, 100% for green, 316% or more for red.

Chapter 7

Conclusion

This chapter reviews and discusses the contributions of this dissertation and presents future work possibilities. The contributions are revisited and discussed in Section 7.1 and a discussion of future research goals and new application areas is presented in Section 7.2. This chapter concludes with a summary in Section 7.3.

7.1 Review of Contributions

This section summarizes the contributions of this dissertation. The contributions are restated in the order presented in the dissertation along with a discussion of how they were accomplished.

1. *A theoretical development showing that minimizing sum-of-Kullback-Leibler divergences, in either ordering of parameters, maximizes a lower bound on Bayes probability of error, a measure of indistinguishability between probability distributions.*

   Multi-class conditional posteriors were the model chosen to represent underlying tissue structure for the multi-modal image set registration and atlas formation. Chapter 2 investigated two methods for producing probability distribution averages via minimizing sum-of-Kullback-Leibler divergences. It was shown that $\hat{p} = \text{argmin}_p \bar{D}\left(\{p_i\}_{i=1}^N || p\right)$ produced the arithmetic mean and that $\hat{p} = \text{argmin}_p \bar{D}\left(p || \{p_i\}_{i=1}^N\right)$ produced the normalized geometric mean. Both averaging methods were determined via the method of Lagrange multipliers.

Using the arithmetic mean, the sum-of-Kullback-Leibler dispersion $\bar{D}\left(\{p_i\}_{i=1}^N||\hat{p}\right)$ was directly related to generalized Jensen-Shannon divergence. Existing Bayes error bounds on Jensen-Shannon divergence presented in [60] therefore provided Bayes error bounds on the population-centric dispersion measure. Minimizing $\bar{D}\left(\{p_i\}_{i=1}^N||\hat{p}\right)$ maximizes a lower bound on Bayes probability of indistinguishability error.

Using the normalized geometric mean, the sum-of-Kullback-Leibler dispersion $\bar{D}\left(\hat{p}||\{p_i\}_{i=1}^N\right)$ was found to provide an upper bound to a dispersion function for interpreting pair-wise Bayes error. Specifically, bounds on Bayes error in terms of the Bhattacharrya coefficient were used to show that minimizing $\bar{D}\left(\hat{p}||\{p_i\}_{i=1}^N\right)$ also maximizes a lower bound on Bayes probability of indistinguishability error.

2. *A novel multi-modal image set registration method is presented. To the author's knowledge this is the only method that incorporates an arbitrary number of multi-modal images per subject. An advantageous consequence of this framework is inverse-invariant (symmetric) registration.*

   A Bayesian framework for generating inter-subject large deformation transformations between two multi-modal sets of images was presented in Chapter 3. Fundamental to this method was the assumption that human brain anatomy consists of finitely enumerable structures. These structures are captured by estimating a class-conditional posterior map for each structure. The modality independent registration framework was achieved by jointly estimating the posterior probabilities associated with the multi-modal image sets and the high-dimensional registration transformations mapping the posteriors. To drive the registration, relative entropy between each of the posteriors and an evolving posterior average, in an independent coordinate space, was minimized. Using the posterior average provided an intrinsically inverse-invariant registration framework. The registration framework was based on the matching problem formulated via fluid flows introduced by [14].

3. *An extension of the above framework to unbiased multi-class atlas formation.*

   The multi-modal image set registration framework was extended to large deformation multi-class posterior atlas estimation in Chapter 4. The method generates

a representative anatomical template from an arbitrary number of topologically similar multi-modal image sets. The definition of the average atlas follows from the notion of Fréchet means. The generated atlas is the class posterior that requires the least amount of deformation energy to be deformed into every class posterior. The method is computationally practical in that computation time grows linearly with the number of image sets. To the author's knowledge this is first unbiased atlas building method that is based on a population of sets of multi-modal images. Constructing such an atlas provides pair-wise correspondence between any two image sets via transformations through said atlas.

4. *The use of information theory to evaluate atlas stability.*

   Entropy of image intensities was used to quantify atlas sharpness in Chapter 5. A blurry image will exhibit greater entropy of intensities than a sharper image. Using this measure, it was shown that resampling an image using linear interpolation adds entropy to the image comparable to that of blurring the image with Gaussian kernel with a width $\sigma = 1$ voxels. For atlas formation, entropy was used to address the question of how many subjects are required to produce a stable atlas. Random permutation tests were conducted to study atlas entropy as a function of the number of images used to produce an atlas. It was shown that, within the choice of interpolation method, the large deformation diffeomorphic atlases were sharp after the inclusion of ten or more subjects. This number is certainly dependent on the particular population used.

5. *An application of the atlas formation to an aging study involving multi-modal brain image data from ninety-five subjects.*

   In Chapter 6, the multi-class posterior atlas formation method was applied to a database of multi-modal images from ninety-five adult brains as part of a healthy aging study. In this study, 4D spatiotemporal atlases were created for the male and female populations. This work is unique in that Fréchet mean atlases were created for a number of time points. These mean atlases suggest an approximation for the line of best fit class posteriors over time. Based on the results in Chapter 5, sliding windows of ten and fifteen subjects were used to ensure temporal smoothness. All of this was facilitated by the good age distribution of the subjects in the database used. Volumetric analysis of white-matter, grey-

matter, and cerebrospinal fluid change over time were consistent with results from previous studies involving large databases.

7.2 Future Work

This section presents several extensions to this dissertation and discusses areas for future research. This section is divided into three sections: Section 7.2.1 describes the creation of continuous 4D spatiotemporal atlases, Section 7.2.2 proposes the extension of multi-modal image set registration to images of subjects with pathology, and Section 7.2.3 describes the application of multi-modal image set registration to multi-center studies were MR images are acquired from scanners of different field-strength.

7.2.1 Continuous 4D Spatiotemporal Atlas

Can regression methods be applied to the 3D spatial atlases at discrete time points? What notion of distance, and subsequently interpolation (particularly temporal), are appropriate in this setting? This dissertation has attempted to provide an approximate solution by moving average windows. The main idea is to overcome the limitations of viewing data as belonging to age groups by approximating a continuous process.

7.2.2 Registration of Images Involving Pathologies

The multi-modal image set registration as presented here might be potentially significant in various applications which rely on the measurement of image sets. For example, multi-modal imaging is standard in the imaging of pathologies such as tumors and lesions. Registration between images presenting pathology and images of healthy subjects is a challenging task since space-occupying lesions have to be treated differently from infiltrating lesions. Specifically, the registration needs to accommodate both local spatial deformation and local change of image intensity. Existing registration method involving scalar images based on image brightness do not accommodate pathologies. In the formation of the class posteriors, one can explicitly assign classes to the various healthy and pathological tissues.

7.2.3 Multi-Center Studies

Another potential application for this method is the registration of images acquired from scanners of different field-strength. Image set registration across different scanners becomes an increasingly important component in multi-center studies. For example, in studies of developmental changes covering multiple years, and in follow-up studies of diseases with change of scanner technology. Images acquired from different scanners potentially have different contrasts and different spatial distortions. The method presented in this dissertation may help address these problems as the registration would be based on underlying anatomical structures rather than simply image intensities. This assumes a robust segmentation method capable of handling the output from scanners of different field strength.

7.3 Summary

This dissertation presented a Bayesian framework for generating large deformation transformations between multi-modal image sets. An image set may be comprised of an arbitrary number of multi-modal images. To the author's knowledge, this is the first such method capable of exploiting the complementary information provided by multi-modal image sets. This modality independent registration framework was achieved by jointly estimating the multi-class posterior probability maps associated with the multi-modal modal image sets and large deformation diffeomorphisms mapping these posterior maps. This framework was extended to large deformation multi-class posterior map atlas estimation. The method generates an unbiased sharp representative anatomical template from an arbitrary number of topologically similar multi-modal image sets. This method was applied to an aging study involving ninety-five subjects to study global and local volumetric change. This research shows promise for future work in building 4D spatiotemporal atlases, image registration involving subjects with pathologies, and multi-center studies.

Appendix A

Information Theoretic Measures

In this appendix, the basic quantities of information theory: entropy, relative entropy, and mutual information, are presented in their discrete form. These measures are functionals of probability distributions and, hence, are not dependent on actual values assumed by random variables. Additionally, a more recent measure, Jensen-Shannon divergence, is also presented. This measure is used in Chapter 2 to define inequalities involving Bayes probability of error P_e. In the medical image analysis context, image intensities can be interpreted as random variables whose behavior can be characterized in terms of probability distributions upon which the basic quantities of information theory can be applied. This dissertation focuses on tissue class-conditional probability mass function maps.

A.1 Entropy

The concept of entropy was first developed in the field of thermodynamics as its second law which states that the entropy of an isolated system is non-decreasing. In searching for a quantity which measures how much, or at what rate, information is produced by a process, Shannon [87], building on the work of Hartley [38], developed the concept of entropy to measure the average uncertainty of a random variable.

Definition A.1 (Uncertainty)**.** *The* uncertainty*, U, of a random variable X is given by*

$$U(x) = -\log p(x)$$

where the probability $p(x)$ is the probability distribution characterizing X .

$U(x)$ is monotonic in $p(x)$ and positive for all values of $p(x)$, since $0 \leq p(x) \leq 1$. Both of these properties are desirable features of a measure. Also, the least and greatest uncertainty occur when $p(x) = 1$ and $p(x) = 0$ respectively. This presents an intuitive understanding for the measure, since if $X = x$ happens with probability 1 then one is certain of X and, hence, have the least uncertainty.

Definition A.2 (Entropy). *The* entropy *of a random variable X, $H(X)$, is defined as the average uncertainty over all possible values X may assume:*

$$\begin{aligned} H(X) &= \mathcal{E}_p[U(x)] \\ &= \mathcal{E}_p[-\log p(x)] \\ &= -\sum_{x \in X} p(x) \log p(x). \end{aligned}$$

Entropy is typically measured in *bits*, *nats*, and *Hartleys* for logarithms of bases two, e, and ten respectively. Unless otherwise stated, entropies stated in this dissertation will be measured in bits. For an axiomatic derivation of entropy see [87].

There are several important properties of entropy noted below:

- *Non-negativity*: $H(X) \geq 0$.
- *Upper bound*: $H(X) \leq \log(N)$. Entropy is maximum when all probabilities are equally likely; equivalently, when the average uncertainty is greatest.
- *Chain rule*: $H(X, Y) = H(X) + H(X|Y)$.
- *Conditioning reduces entropy*: $H(X|Y) \leq H(X)$.
- *Concavity*: $H(\{p(x_i)\})$ is concave in p.

The notion of entropy can be extended to multiple random variables as shown below.

Definition A.3 (Joint Entropy). *The* joint entropy*, $H(X, Y)$, of a pair of discrete random variables X and Y with joint distribution $p(x, y)$ is defined by*

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y).$$

Definition A.4 (Conditional Entropy). *Let X and Y be discrete random variables with joint distribution $p(x, y)$ and conditional distribution $p(x|y)$. Then the entropy conditioned on a single event is defined by*

$$H(X|Y = y) = -\sum_{x \in X} p(x|y) \log p(x|y).$$

The conditional entropy, *or* equivocation, *is then defined by*

$$\begin{aligned} H(X|Y) &= \sum_{y \in Y} p(y) H(X|Y=y) \\ &= -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log p(x|y) \\ &= -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x|y). \end{aligned}$$

A.2 Kullback-Leibler Divergence

With this well-defined notion of entropy, Kullback and Leibler [58] define a distance measure between two distributions.

Definition A.5 (Kullback-Leibler Divergence)**.** *The* Kullback-Leibler divergence *(or relative entropy) between two probability distributions p and q over the same discrete random variable X is defined by the expected logarithm of likelihood ratio of p to q*

$$\begin{aligned} D(p||q) &= \mathcal{E}_p \left[\log \frac{p(x)}{q(x)} \right] \\ &= \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}. \end{aligned}$$

In a signal model, this measure can be interpreted as the inefficiency of assuming that q is true when p is true. That is, given a model expressed as a probability distribution p, one can then measure how far an observation, also expressed by a probability distribution, q, deviates from p using relative entropy. Additionally, $D(p||q)$ can be viewed as the average number of bits that are wasted by encoding events from distribution p with a code based on distribution q.

The Information Inequality theorem provides the basic properties of Kullback-Leibler divergence.

Theorem A.1 (Information Inequality)**.** *Let $p(x)$ and $q(x)$ be two probability mass functions associated with random variable X. Then $D_{KL}(p||q) \geq 0$ with equality if and only if $p(x) = q(x)$ for all $x \in X$.*

Proof. See [16], page 26, for an argument based on Jensen's Inequality and fact that the function $-\log$ is convex. □

Although a useful distance function, Kullback-Leibler divergence is not a metric. By Theorem A.1, $D(\cdot||\cdot)$ satisfies the non-negativity and identity of indiscernables properties of a metric, Definition 1.1. $D(\cdot||\cdot)$, however, is not a metric since it does not obey the symmetry and triangle inequality properties. Consider the following simple counterexample.

Example A.1 (Kullback-Leibler divergence is not a metric.)**.** *Consider the following three probability mass functions,*

$$p(x) = \left\{\frac{1}{2}, \frac{1}{2}\right\} \quad q(x) = \left\{\frac{3}{4}, \frac{1}{4}\right\} \quad r(x) = \left\{\frac{7}{8}, \frac{1}{8}\right\}.$$

In computing Kullback-Leibler divergences between these probability mass functions, note that

$$\begin{aligned} D(p||q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\ &= \frac{1}{2} \log \left(\frac{1}{2} \cdot \frac{4}{3}\right) + \frac{1}{2} \log \left(\frac{1}{2} \cdot \frac{4}{1}\right) \\ &= \frac{1}{2} \log \left(\frac{2}{3}\right) + \frac{1}{2} \log 2 \\ &= \frac{1}{2} - \frac{1}{2} \log 3 + \frac{1}{2} \\ &= 1 - \frac{1}{2} \log 3 \end{aligned}$$

and

$$\begin{aligned} D(q||p) &= \sum_x q(x) \log \frac{q(x)}{p(x)} \\ &= \frac{3}{4} \log \left(\frac{3}{4} \cdot \frac{2}{1}\right) + \frac{1}{4} \log \left(\frac{1}{4} \cdot \frac{2}{1}\right) \\ &= \frac{3}{4} \log \frac{3}{2} + \frac{1}{4} \log \frac{1}{2} \\ &= \frac{3}{4} \log 3 - \frac{3}{4} - \frac{1}{4} \\ &= \frac{3}{4} \log 3 - 1. \end{aligned}$$

That is, $D(p||q) > D(q||p)$, and, hence, $D(\cdot||\cdot)$ is not symmetric. In computing two

more Kullback-Leibler divergences, one finds that

$$\begin{aligned} D(q||r) &= \sum_x q(x) \log \frac{r(x)}{q(x)} \\ &= \frac{3}{4} \log \left(\frac{3}{4} \cdot \frac{8}{7} \right) + \frac{1}{4} \log \left(\frac{1}{4} \cdot \frac{8}{1} \right) \\ &= \frac{3}{4} \log \frac{6}{7} + \frac{1}{4} \log 2 \\ &= \frac{3}{4} \log 2 + \frac{3}{4} \log 3 - \frac{3}{4} \log 7 + \frac{1}{4} \\ &= 1 + \frac{3}{4} \log 3 - \frac{3}{4} \log 7 \end{aligned}$$

and

$$\begin{aligned} D(p||r) &= \sum_x p(x) \log \frac{p(x)}{r(x)} \\ &= \frac{1}{2} \log \left(\frac{1}{2} \cdot \frac{8}{7} \right) + \frac{1}{2} \log \left(\frac{1}{2} \cdot \frac{8}{1} \right) \\ &= \frac{1}{2} \log \frac{4}{7} + \frac{1}{2} \log 4 \\ &= \frac{1}{2} \log 4 - \frac{1}{2} \log 7 + 1 \\ &= 2 - \frac{1}{2} \log 7. \end{aligned}$$

Now consider the expression

$$\begin{aligned} D(p||q) + D(q||r) - D(p||r) &= 1 - \frac{1}{2} \log 3 + 1 + \frac{3}{4} \log 3 - \frac{3}{4} \log 7 - 2 + \frac{1}{2} \log 7 \\ &= \frac{1}{4} \left(\log 3 - \log 7 \right) \\ &< 0 \\ \Rightarrow D(p||q) + D(q||r) &< D(p||r). \end{aligned}$$

That is, $D(\cdot||\cdot)$ does not obey the triangle inequality.

A.3 Jensen-Shannon Divergence

A generalized notion of Kullback-Leibler divergence, Jensen-Shannon divergence, will be used in Chapter 2 to provide bounds on Bayes probability of error. The

Kullback-Leibler divergence is always non-negative, but can be unbounded (e.g., when $p_1(x) \neq 0$ and $p_2(x)$, $D(p_1||p_2) = \infty$) and is, as noted above, not symmetric. A more recent measure between two probability distributions is the Jensen-Shannon divergence introduced by [60] which is both bounded and symmetric.

Definition A.6 (Jensen-Shannon Divergence)**.** *Let* $\pi = \{\pi_1, \pi_2\}$ *with* $\pi_1, \pi_2 \geq 0$ *and* $\pi_1 + \pi_2 = 1$ *be prior probabilities on two probability distributions* $p_1(x)$ *and* $p_2(x)$*. The* Jensen-Shannon divergence *between* p_1 *and* p_2 *defined by*

$$\begin{aligned} JS_\pi(p_1||p_2) &= H(\pi_1 p_1 + \pi_2 p_2) - \pi_1 H(p_1) - \pi_2 H(p_2) \\ &= \pi_1 D(p_1||M_\pi) + \pi_2 D(p_2||M_\pi) \end{aligned}$$

where $M_\pi = \pi_1 p_1 + \pi_2 p_2$ *is called the* mutual source *of* p_1 *and* p_2 [23].

Definition A.7 (Generalized Jensen-Shannon Divergence)**.** *The Jensen-Shannon divergence can be generalized to measure the distance between any finite number of probability distributions as*

$$\begin{aligned} JS_\pi(\{p_i\}_{i=1}^N) &= H\left(\sum_{i=1}^N \pi_i p_i\right) - \sum_{i=1}^N \pi_i H(p_i) \\ &= -\sum_{x\in X}\left(\sum_{i=1}^N \pi_i p_i(x)\right)\log\left(\sum_{j=1}^N \pi_j p_j(x)\right) + \sum_{i=1}^N \pi_i \sum_{x\in X} p_i(x)\log p_i(x) \\ &= \sum_{i=1}^N \sum_{x\in X} \pi_i p_i(x)\left[-\log\left(\sum_{j=1}^N \pi_j p_j(x)\right) + \log p_i(x)\right] \\ &= \sum_{i=1}^N \sum_{x\in X} \pi_i p_i(x)\log\frac{p_i(x)}{\sum_{j=1}^N \pi_j p_j(x)} \\ &= \sum_{i=1}^N \pi_i D(p_i||M_\pi) \end{aligned}$$

where $M_\pi = \sum_{j=1}^N \pi_j p_j$ *with* $\pi_i \geq 0$ *and* $\sum_{i=1}^N \pi_i = 1$.

Although not used in the framework proposed by this dissertation, for completeness, this section concludes with the definition of the last of three fundamental information theoretic measures, mutual information. Mutual information will be discussed later in the context of related work.

A.4 Mutual Information

The last of the three basic information theoretic measures is mutual information: the measure of the amount of information one random variable X has about another Y.

Definition A.8 (Mutual Information)**.** Mutual information *is defined as the relative entropy between the joint distribution and the product of two probability mass functions:*

$$\begin{aligned} I(X;Y) &= D(p(x,y)||p(x)p(y)) \\ &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}. \end{aligned}$$

Mutual information measures the dependence of random variables X and Y. When X and Y are independent $p(x,y) = p(x)p(y)$ and $I(X;Y) = 0$. Mutual information can also be expressed in terms of entropy as follows:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y). \end{aligned} \tag{A.1}$$

In equation A.1, mutual information is viewed as the reduction in the uncertainty of X due to the knowledge of Y. There are several important properties of mutual information noted below:

- *Non-negativity*: $I(X;Y) \geq 0$.
- *Symmetry*: $I(X;Y) = I(Y;X)$.
- *Self-information*: $I(X;X) = H(X)$.
- *Independence*: $I(X;Y) = 0 \iff X \perp Y$.
- *Information explanation*:$I(X;Y) \leq H(X)$ and $I(Y;X) \leq H(Y)$.

A.5 Multivariate Mutual Information

The two notions of mutual information in Section A.4 can be extended to the multivariate setting.

A.5.1 Entropy Reduction

The first extension follows the "entropy reduction" idea espoused by [1] which is a generalization of Equation A.1,

$$I(X_1;\ldots;X_N) = \sum_{k=1}^{N} \sum_{\{i_1,\ldots,i_k\}\subseteq\{1,\ldots,N\}} (-1)^{k+1} H(p_{i_1},\ldots,p_{i_k}).$$

In this context, multivariate mutual information amount $N \geq 2$ random variables can be interpreted as the measure of their simultaneous interaction [105, 56]. It can be equivalently interpreted as a multi-way similarity measure among random variables. If the value is zero, the N random variables do not simultaneously interact. This generalization of mutual information can result in negative values. As an example, consider the three random variable case, $\{X_i\}_{i=1}^{3}$,

$$\begin{aligned} I(X_1,X_2,X_3) = {} & H(X_1) + H(X_2) + H(X_3) \\ & - H(X_1,X_2) - H(X_1,X_3) - H(X_2,X_3) \\ & + H(X_1,X_2,X_3). \end{aligned}$$

Note that the symmetry property still holds.

A.5.2 Redundancy Measure

The generalization of mutual information based on Definition A.8 preserves the non-negativity property. This is the approach of [105] where they extend the relative entropy concept to a "redundancy measure"

$$\begin{aligned} R(X_1,\ldots,X_N) &= D\left(p(x_1,\ldots,x_N)||\prod_{k=1}^{N} p(x_k)\right) \\ &= \sum_{k=1}^{N} H\left(p(x_k)\right) - H\left(p(x_1,\ldots,x_N)\right). \end{aligned}$$

By the Information Inequality Theorem, TheoremA.1, this measure is non-negative and equal to zero if the X_i are stochastically independent. The higher the redundancy among the random variables, the stronger their functional dependency [46, 47]

Appendix B

Study of Convergence

B.1 Introduction

How likely is it that the actual optimum in the multi-modal image set registration cost function is achieved? It is impossible to know for real data, since there is no way to know the true correspondence. For synthetic transformations and synthetic image intensities, the algorithm comes close to the global minimum of the cost function. Using synthetic data, this document presents a qualitative evaluation of the registration and a quantitative evaluation of the data likelihood estimation under the multi-variate normal distribution model.

B.2 Synthetic Data Example[1]

To evaluate the performance of the algorithm, a synthetic 2D dataset was generated. Specifically, a geometric prior, a known transformation, and two synthetic images whose radiometric characteristics are statistically similar to actual T1- and T2-weighted MR images were defined. A four-class atlas prior comprised of concentric ellipses was generated using Matlab. The subsequent composite label image was generated by the superposition of the individual classes. Both the atlas and composite label image are shown in Figure B.1. Additionally, a transformation was constructed using sinusoidal displacements, which was then applied to the composite label image to produce the foundation for the synthetic image set. The multi-modal synthetic image set was created from two images that were simulated by sampling from a multi-variate Gaussian distribution with different means and covariances for each of the classes in the deformed label image. The deformed label image and the corresponding multi-modal images are shown in Figure B.2. The algorithm was run for fifty iterations with ten steps of the large deformation diffeomorphic registration per iteration. The final segmentation and deformation estimates are also shown in Figure B.3.

[1] This section represents portions of the WBIR 2003 paper [62].

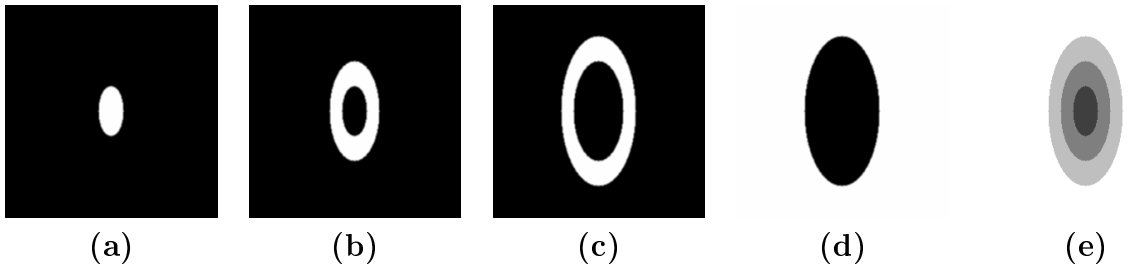


Figure B.1: Geometric Atlas Prior

The manually generated geometric four-class atlas prior (a-d) and the corresponding composite labeled image (e).

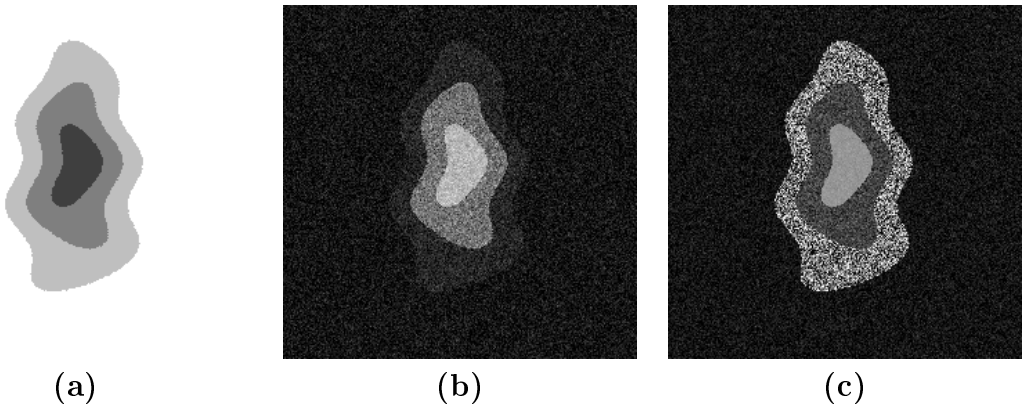


Figure B.2: Synthetic Image Set

The deformed label image (a), the synthetic image derived from T1 samples (b), and the synthetic image derived from T2 samples (c).

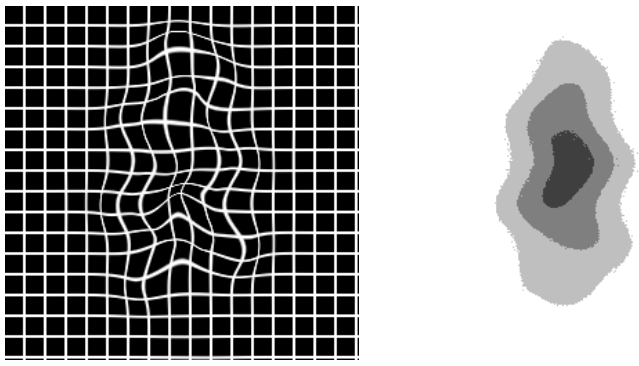


Figure B.3: Estimated Deformation

The estimated deformation expressed as a warped regular grid (a) and the estimated segmentation (b) following fifty iterations of the algorithm.

To evaluate the registration the algorithm was run, again with fifty iterations, holding the transformation fixed to the identity map. With the transformation fixed to the identity map, the expectation maximization provided the maximum likelihood solution. The final estimated segmentation was then compared to the registration-based segmentation, with results shown in Figure B.4. By examining the regions where the two segmentations differ from the ground truth label image, it is clear the registration has improved the segmentation.

In both invocations of the algorithm, the class means and covariances were collected and compared. Figure B.5 shows the final relative norms for the estimated and actual means and covariances at the final iteration. For all classes, the registration has improved estimates for both the means and covariances. The convergence of the mean and covariance estimates using registration is shown in Figure B.6. This figure shows that the estimates of the means and covariances have converged quickly when the transformation is fixed to the identity map. When registration is added, the estimates of the means and covariances continue to improve as the estimation of the transformation between the atlas and the subject converges. This exemplifies the effectiveness of the alternating nature of the algorithm. These results show that the registration improves the segmentation by accommodating local variability.

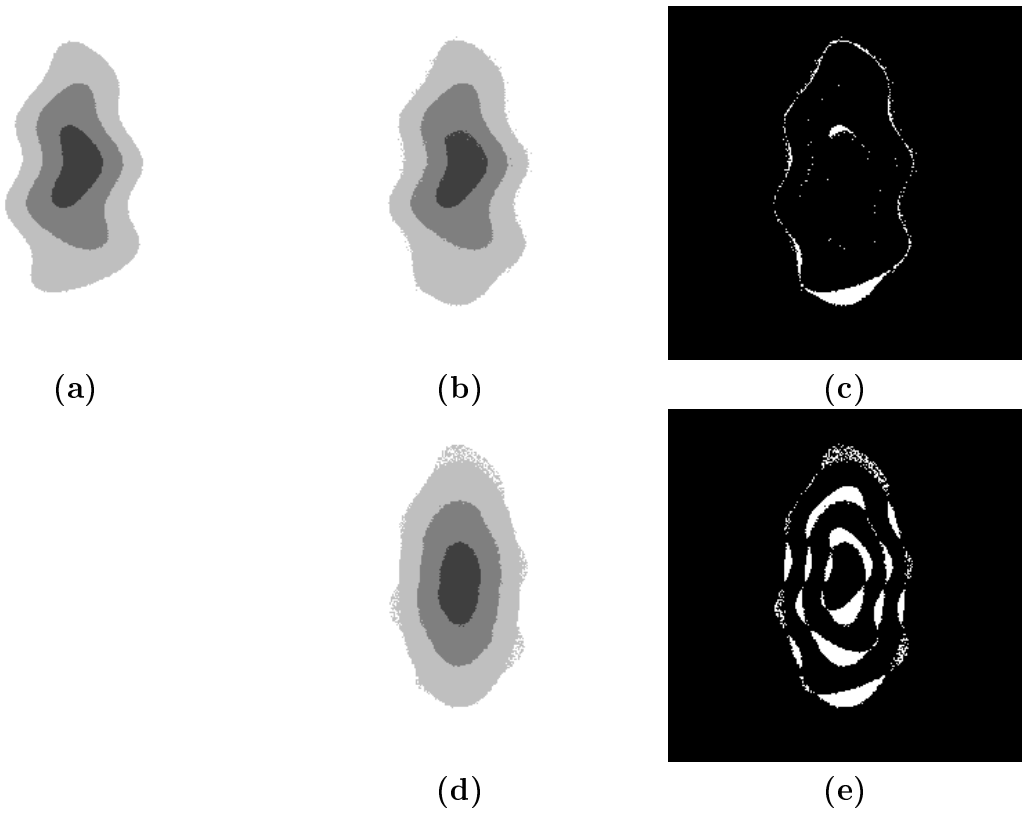


Figure B.4: Final Segmentation

The top row shows the ground truth label image (a), the final segmentation estimation using registration (b), and the regions where this segmentation differs from the ground truth (c). The bottom row shows the final segmentation estimation without using registration (d), and regions where this segmentation differs from the ground truth (e).

Class	$\frac{\|\hat{\mu}_{c_i}-\mu_{c_i}\|}{\|\mu_{c_i}\|}$w/ reg.	$\frac{\|\hat{\mu}_{c_i}-\mu_{c_i}\|}{\|\mu_{c_i}\|}$w/o reg.	$\frac{\|\hat{\Sigma}_{c_i}-\Sigma_{c_i}\|_F}{\|\Sigma_{c_i}\|_F}$w/ reg.	$\frac{\|\hat{\Sigma}_{c_i}-\Sigma_{c_i}\|_F}{\|\Sigma_{c_i}\|_F}$w/o reg.
c_1	0.0047	0.1216	0.0450	7.0588
c_2	0.0152	0.1168	0.2104	1.8164
c_3	0.0960	0.0939	0.0206	0.0811
c_4	0.0046	0.0046	0.0081	0.0266

Figure B.5: Relative Norm Statistics

The first two columns of numbers are the means at the final iteration $\hat{\mu}_{c_i}$ relative to the actual means μ_{c_i} using registration and fixed identity map. The last two columns show the same for the relative covariances using the Frobenius norm, $||A||_F = \sqrt{\mathrm{tr}\,(AA^T)}$.

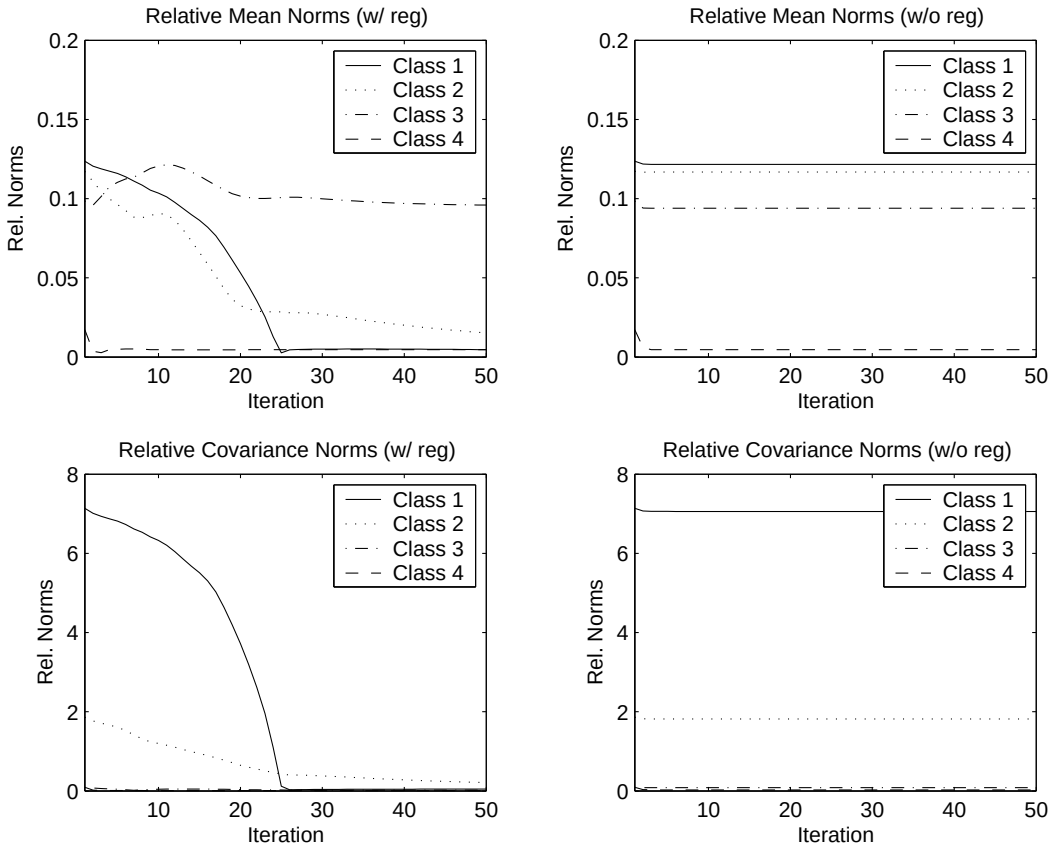


Figure B.6: Convergence

The left column shows the convergence of means and covariances using registration. The right column shows the same using the fixed identity transformation.

Bibliography

[1] Norm Abramson. *Information Theory and Coding.* McGraw-Hill, 1963.

[2] J. Andersson and L. Thurfjell. A multivariate approach to registration of dissimilar tomographic images. *European Journal of Nuclear Medicine*, 26:718–733, 1999.

[3] J. Ashburner and K. J. Friston. Unified segmentation. *NeuroImage*, 26(3):839–851, April 2004.

[4] David Atkinson, Derek L. G. Hill, Peter N. R. Stoyle, Paul E. Summers, and Stephen F. Keevil. Automatic correction of motion artifacts in magnetic resonance images using an entropy focus criterion. *IEEE Transactions on Medical Imaging (TMI)*, 16(6):903–910, December 1997.

[5] Brain Avants and James C. Gee. Geodesic estimation for large deformation anatomical shape averaging and interpolation. *NeuroImage*, Supplement 1:S139–S150, 2004.

[6] Ravi Bansal, Lawrence H. Staib, Zhe Chen, Anand Rangarajan, Jonathan Knisely, Ravinder Nath, and James S. Duncan. Entropy-based, dual-portal-to-3DCT registration incorporating pixel correlation. *IEEE Transactions on Medical Imaging (TMI)*, 22(1):29–49, January 2003.

[7] Mirza Faisal Beg. *Variational and Computational Methods for Flows of Diffeomorphisms in Image Matching and Growth in Computational Anatomy.* PhD thesis, The Johns Hopkins University, August 2003.

[8] J. Besag. On the statistical analysis of dirty pictures. *Journal of Royal Statistical Society, Series B*, 48:259–302, 1986.

[9] K. K. Bhatia, J. V. Hajnal, B. K. Puri, A. D. Edwards, and D. Ruekert. Consistent groupwise non-rigid registration for atlas construction. In *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 908–911, April 2004.

[10] Fred L. Bookstein. *Morphometric Tools for Landmark Data.* Cambridge University Press, 1991.

[11] K Brodmann. *Vergleichende Lokalisationslehre der Grosshimrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues.* Barth, Leipzig, 1909.

[12] Pascal Cachier and David Rey. Symmetrization of the non-rigid registration problem using inversion-invariant energies: Application to multiple sclerosis. In *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Lecture Notes in Computer Science (LNCS), pages 472–481. Springer-Verlag, 2002.

[13] Ho-Ming Chan, Albert C. S. Chung, Simon C. H. Yu, Alexander Norbash, and William M. Wells III. Multi-modal image registration by minimizing Kullback-Leibler distance between expected and observed joint class histograms. In *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 570–576. IEEE, 2003.

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

[15] Andre Collignon, Frederick Maes, Dirk Vandermeulen, G. Marchal, and Paul Suetens. Automated multimodality image registration using information theory. *Proceedings of Information Processing in Medical Imaging (IPMI)*, pages 263–274, 1995.

[16] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and Sons, Inc., New York, 1991.

[17] Mathieu De Creane, Aloy du Bois d'Aische, Benoît Macq, and Simon K. Warfield. Multi-subject registration for unbiased statistical atlas construction. In *Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Lecture Notes in Computer Science (LNCS), pages 655–662. Springer-Verlag, September 2004.

[18] JG Csernansky, S Joshi, LE Wang, J Haller, M Gado, MI Miller, and U Grenander. Hippocampal morphometry in schizophrenia via high dimensional brain map ping. In *Proceedings of the National Academy of Sciences, USA*, volume 95, pages 11406–11411, September 1998.

[19] J.G. Csernansky, L. Wang, J. Swank, J. P. Miller, M. Gado, D. McKeel, M. I. Miller, and J. C. Morris. Preclinical detection of alzheimer's disease: hippocampal shape and volume predict dementia onset in the elderly. *NeuroImage*, 25:783–792, 2005.

[20] E. D'Agostino, F. Maes, D. Vandermeulen, and P. Suetens. An information theoretic approach for non-rigid image registration using voxel class probabilities. *to appear in Medical Image Analysis (MedIA)*, 2006.

[21] Richard Duda, Peter Hart, and Richard Stork. *Pattern Classification*. John Wiley and Sons, Inc., New York, 2000.

[22] Paul Dupuis, Ulf Grenander, and Michael Miller. Variational problems on flows of diffeomorphisms for image matching. *Quarterly Journal of Applied Mathematics*, pages 587–600, November 1998.

[23] Ran El-Yaniv, Shai Fine, and Naftali Tishby. Agnostic classification of Markovian sequences. In *Advances in Neural Information Processing Systems*, volume 10. The MIT Press, 1998.

[24] A .C. Evans, D. L. Collins, S. R. Mills, E. D. Brown, R. L. Kelley, and T. M. Peters. 3D statistical neuroanatomical models from 305 MRI volumes. In *IEEE Nuclear Science Symposium and Medical Imaging Conference*, pages 1813–1817, 1993.

[25] F. Fazekas, F. Barkhof, M. Filippi, R. I. Grossman, D. K. Li, W. I. McDonald, H. F. McFarland, D. W. Paty, J. H. Simon, J. S. Wolinsky, and D. H. Miller. The contribution of magnetic resonance imaging to the diagnosis of multiple sclerosis. *Neurology*, 53(3):448–456, August 1999.

[26] P. Thomas Fletcher, Conglin Lu, Stephen M. Pizer, and Sarang C. Joshi. Principal geodesic analysis for the study of nonlinear statistics in shape. *IEEE Transactions on Medical Imaging (TMI)*, 23(8):995–1005, 2004.

[27] Maurice Fréchet. Les elements aleatoires de nature quelconque dans un espace distancie. *Annales De L'Institut Henri Poincare*, 10:215–310, 1948.

[28] Peter A. Freeboro and Nick C. Fox. Modeling brain deformations in alzheimer disease by fluid registration of serial 3D MR images. *Journal of Computer Assisted Tomography*, 22(5):838–843, September 1998.

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

[30] Catriona D. Good, Ingrid S. Johnsrude, John Ashburner, Richard N. A. Henson, Karl J. Friston, and Richard S. J. Frackowiak. A voxel-based morphometric study of ageing in 465 normal adult human brains. *NeuroImage*, 14:21–36, 2001.

[31] Ulf Grenander. *General Pattern Theory*. Oxford University Press, 1993.

[32] Ulf Grenander and Michael Miller. Computational anatomy: an emerging discipline. *Quarterly of Applied Mathematics*, 56:617–694, 1998.

[33] Régis Guillemaud and Michael Brady. Estimating the bias field of MR images. *IEEE Transactions on Medical Imaging (TMI)*, 16(6):878–886, 1997.

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

[35] R. C. Gur, F. M. Gunning-Dixon, B. I. Turetsky, W. B. Bilker, and R. E. Gur. Brain region and sex differences in age association with brain volume: a quantitative MRI study of healthy young adults. *American Journal of Geriatric Psychiatry*, 10:72–80, 2002.

[36] C. R. Guttmann, F. A. Jolesz, R. Kikinis, R. J. Killiany, M. B. Moss, T. Sandor, and M. S. Albert. White matter change with normal aging. *Neurology*, 50(4):972–978, 1998.

[37] G. H. Hardy, J. E. Littlewood, and G. Polya. *Inequalities*. The Syndics of the Cambridge University Press, 1964.

[38] R. V. L. Hartley. Transmission of information. *Bell System Technical Journal*, 7:535–563, 1928.

[39] Jianchun He and Gary E. Christensen. Large deformation inverse consistent elastic image registration. In *Information Processing in Medical Imaging (IPMI)*, Lecture Notes in Computer Science (LNCS), pages 438–449. Springer-Verlag, 2003.

[40] Y. He, A. Hamza, and H. Krim. A generalized divergence measure for robust image registration. *IEEE Transactions on Signal Processing*, 51(5), 2003 2003.

[41] Martin E. Hellman and Josef Raviv. Probability of error, equivocation, and the chernoff bound. *IEEE Transactions on Information Theory*, 16(4):368–372, July 1970.

[42] Gerardo Hermosillo. *Variational Methods for Multimodal Image Matching*. PhD thesis, Universite de Nice - Sophia Antipolis, May 2002.

[43] Sean Ho, Elizabeth Bullitt, and Guido Gerig. Level set evolution with region competition: automatic 3-D segmentation of brain tumors. In *Proceedings of the 16th International Conference on Pattern Recognition*, pages 532–535. IEEE Computer Society, August 2002.

[44] K. H. Höhne, M. Bomans, M. Riemer, U. Tiede, R. Shubert, and W. Lierse. A 3D anatomical atlas based on a volume model. *IEEE Computer Graphics Applications*, December 1992.

[45] Inseok Hwang, Kaushik Roy, Hamsa Balakrishnan, and Claire Tomlin. A distributed multiple-target identity management algorithm in sensor networks. In *IEEE Conference on Decision and Control*, December 2004.

[46] H. Joe. Estimation of entropy measures of multivariate dependence. *Annals of the Institute of Statistical Mathematics*, 41:683–697, 1989.

[47] H. Joe. Relative entropy measures of multivariate dependence. *Journal of American Statistical Association*, 48:157–164, 1989.

[48] S. Joshi and M. I. Miller. Landmark matching via large deformation diffeomorphisms. *IEEE Transactions on Image Processing (TIP)*, 9(8):1357–1370, August 2000.

[49] Sarang Joshi, Brad Davis, Matthieu Jomier, and Guido Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage: Supplement issue on Mathematics in Brain Imaging*, 23(Supplement1):S151–S160, 2004.

[50] Sarang Joshi, Peter Lorenzen, Guido Gerig, and Elizabeth Bullitt. Structural and radiometric asymmetry in brain images. *Medical Image Analysis (MedIA)*, 7(2):155–170, June 2003.

[51] Sarang C. Joshi, Ulf Grenander, and Michael I. Miller. On the geometry of and shape of brain sub-manifolds. *International Journal of Pattern Recognition and Artificial Intelligence: Special Issue on Processing of MR Images of the Human*, 11(8):1317–1343, 1997.

[52] Sarang C. Joshi, Michael I. Miller, Gary E. Christensen, Ayananshu Banerjee, Thomas A. Coogan, and Ulf Grenander. Hierarchical brain mapping via a generalized dirichlet solution for mapping brain manifolds. In *Proc. of the SPIE's 1995 International Symposium on Optical Science, Engineering, and Instrumentation*, volume 2573, pages 278–289, August, 1995.

[53] T. Kailath. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Transactions on Communication Theory*, (1):52–60, February 1967.

[54] D. G. Kendall. Shape manifolds, procrustean metrics and complex projective spaces. *Bulletin of the London Mathematical Society*, 16:81–121, 1984.

[55] Ron Kikinis, Martha Shenton, Dan Iosifescu, Robert McCarley, Pai Saiviroonporn, Hiroto Hokama, Andre Robatino, David Metcalf, Cynthia Wible, Chiara Portas, Robert Donnino, and Ferenc A. Jolesz. A digital brain atlas for surgical planning, model driven segmentation and teaching. *IEEE Transactions on Visualization and Computer Graphics*, 2(3):232–241, September 1996.

[56] I. Kojadinovic. Agglomerative hierarchial clustering of continuous variables based on mutual information. *Computational Statistics and Data Analysis*, 46:269–294, 2004.

[57] Natasa Kovacevic, Josette Chen, John G. Sled, Jeff Henderson, and Mark Henkelman. Deformation based representation of groupwise average and variability. In *Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Lecture Notes in Computer Science (LNCS), pages 615–622. Springer-Verlag, September 2004.

[58] Soloman Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, March 1951.

[59] D. N. Levin, A. Herrmann, T. Spraggins, P. A. Collins, L. B. Dixon, M. A. Simon, and A. E. Stillman. Musculoskeletal tumors: improved depiction with linear combinations of MR images. *Radiology*, 163:545–549, 1987.

[60] Jianhua Lin. Divergence measures based on shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, January 1991.

[61] Peter Lorenzen, Brad Davis, and Sarang Joshi. Unbiased atlas formation via large deformations metric mapping. In Jim Duncan and Guido Gerig, editors, *Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Lecture Notes in Computer Science (LNCS), pages 411–418. Springer-Verlag, October 2005.

[62] Peter Lorenzen and Sarang Joshi. High-dimensional multi-modal image registration. In *International Workshop on Biomedical Image Registration (WBIR)*, volume 2717 of *Lecture Notes in Computer Science (LNCS)*, pages 234–243. Springer-Verlag, June 2003.

[63] Peter Lorenzen, Marcel Prastawa, Brad Davis, Guido Gerig, Elizabeth Bullitt, and Sarang Joshi. Multi-modal image set registration and atlas formation. *to appear in Medical Image Analysis (MedIA)*, 2006.

[64] G. D. Tourassi M. P. Wachowiak, R. Smolikova and A. S. Elmaghraby. Generalized mutual information similarity metrics for multimodal biomedical image registration. In *Second Joint EMBS/BMES Conference*, pages 1005–1006, October 2002.

[65] G. D. Tourassi M. P. Wachowiak, R. Smolikova and A. S. Elmaghraby. Similarity metrics based on nonadditive entropies for 2d-3d multimodal biomedical image registration. In *International Society for Optical Engineering (SPIE)*, pages 1090–1100, 2003.

[66] V. A. Magnotta, H. J. Bockholt, H. F. Johnson, G. E. Christensen, and N. C. Andreaseb. Subcortical, cerebellar, and magnetic resonance based consistent brain image registration. *NeuroImage*, 19:233–245, June 2003.

[67] M. Matsumae, R. Kikinis, I. Mórocz, A. Lorenzo, T. Sándor, M. S. Albert, P. McL. Black, and F. Jolesz. Age related changes in intracranial compartment volumes in normal adults assessed by mri. *Journal of Neurosurgery*, 84(6):982–991, 1996.

[68] M. I. Miller and L. Younes. Group actions, homeomorphisms, and matching: A general framework. *Internation Journal of Computer Vision (IJCV)*, 41(1-2):61–84, 2001.

[69] Michael Miller, Ayananshu Banerjee, Gary Christensen, Sarang Joshi, Navin Khaneja, Ulf Grenander, and Larissa Matejic. Statistical methods in computational anatomy. *Statistical Methods in Medical Research*, 6:267–299, 1997.

[70] Michael I. Miller, Sarang C. Joshi, and Gary Christensen. Large deformation fluid diffeomorphism for landmark and image matching. In *Brain Warping*, chapter 7, pages 115–131. Academic Press, San Diego, 1999.

[71] Michael I. Miller, Alain Trouve, and Laurent Younes. On the metrics and euler-lagrange equations of computational anatomy. *Annual Review of Biomedical Engineering*, 4:375–405, August 2002.

[72] Mark Moelich and Tony Chan. Joint segmentation and registration using logic models. CAM 03-06, UCLA, 2003.

[73] Bénédicte Mortamet, Donglin Zeng, Guido Gerig, Marcel Prastawa, and Elizabeth Bullitt. Effects of healthy aging measured by intracranial compartment volumes using a designed MR brain database. In Jim Duncan and Guido Gerig, editors, *Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Lecture Notes in Computer Science (LNCS), pages 383–391. Springer-Verlag, October 2005.

[74] Huzefa Neemuchwala, Alfred Hero, and Paul Carson. Image registration using alpha-entropy measures and entropic graphs. *European Journal on Signal Processing*, 2002.

[75] Frank H. Netter. *Atlas of Human Anatomy*. Novartis, 1997.

[76] Josien Pluim, Antoine Maintz, and Max Viergever. Mutual information based registration of medical images: a survey. *IEEE Transactions on Medical Imaging (TMI)*, 22(8):986–1004, August 2003.

[77] Josien Pluim, Antoine Maintz, and Max Viergever. f-information measures in medical image registration. *IEEE Transactions on Medical Imaging (TMI)*, 23(12):1508–1516, December 2004.

[78] M. Prastawa, J. H. Gilmore, W. Lin, and G. Gerig. Automatic segmentation of MR images of the developing newborn brain. *Medical Image Analysis (MedIA)*, 9:457–466, October 2005.

[79] Marcel Prastawa, Elizabeth Bullitt, Sean Ho, and Guido Gerig. A brain tumor segmentation framework based on outlier dection. *Medical Image Analysis (MedIA)*, pages 275–283, September 2004.

[80] Marcel Prastawa, Nathan Moon, Elizabeth Bullitt, Koen van Leemput, and Guido Gerig. Automatic brain tumor segmentation by subject specific modification of atlas priors. *Academic Radiology*, 10:1341–1348, December 2003.

[81] Alfred Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Math. Stat. Prob.*, volume I, page 547, Berkeley, 1961. University of California Press.

[82] S. M. Resnick, A. F. Goldszal, C. Davatzikos, S. Golski, M. A. Kraut, E. J. Metter, R. N. Bryan, and A. B. Zonderman. One-year age changes in MRI brain volumes in older adults. *Cerebral Cortex*, 10(5):464–472, 2000.

[83] Johannes W. Rohen and Chihiro Yokochi. *Color Atlas of Anatomy: A Photographic Study of the Human Body*. Igaku-Shoin, 1993.

[84] Torsten Rohlfing, Daniel B. Russakoff, and Jr. Calvin R. Mauerer. Extraction and application of expert priors to combine multiple segmentations of human brain tissue. In *Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 2732, pages 578–585. Springer-Verlag, 2003.

[85] Daniel B. Russakoff, Carlo Tomasi, Torsten Rohlfing, and Jr. Calvin R. Mauerer. Image similarity using mutual information of regions. *European Conference on Computer Vision (ECCV)*, pages 596–607, 2004.

[86] J. A. Schnabel, D. Rueckert, M. Quist, J. M. Blackall, A. D. Castellano Smith, T. Hartkens, G. P. Penney, W. A. Hall, H. Liu, C. L. Truwit, F. A. Gerritsen, D. L. G. Hill, and D. J. Hawkes. A generic framework for non-rigid registration based on non-uniform multi-level free-form deformations. In *Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Lecture Notes in Computer Science (LNCS), pages 573–581. Springer-Verlag, October 2001.

[87] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423,623–656, July 1948.

[88] Dinggang Shen and Christos Davatzikos. HAMMER: hierarchical attribute matching mechanism for elastic registration. *IEEE Transactions on Medical Imaging (TMI)*, 21(11):1421–1439, November 2002.

[89] Dinggang Shen and Christos Davatzikos. Measuring temporal morphological changes robustly in brain MR images via 4-dimensional template warping. *NeuroImage*, 21:1508–1517, April 2004.

[90] Colin Studholme. Simultaneous population based image alignment for template free spatial normalization of brain anatomy. In *International Workshop on Biomedical Image Registration (WBIR)*, volume 2717 of *Lecture Notes in Computer Science (LNCS)*, pages 81–90. Springer-Verlag, June 2003.

[91] Colin Studholme, John A. Little, Graeme P. Penny, Derek L. G. Hill, and David J. Hawkes. Automated multimodality registration using the full affine transformation: Application to MR and CT guided skull based surgery. In

Karl Heinz Höhne and Ron Kikinis, editors, *Visualization in Biomedical Computing*, Lecture Notes in Computer Science (LNCS), pages 601–606. Springer-Verlag, 1996.

[92] Jean Talairach and Pierre Tournoux. *Co-Planar Stereotaxis Atlas of the Human Brain.* Georg Thiem Verlag, 1988.

[93] D'Arcy Thompson. *On Growth and Form.* Cambridge University Press, 1917.

[94] Paul Thompson, R. Woods, M. Mega, and Arthur Toga. Mathmatical/computational challenges in creating deformable and probabilistic atlases of the human brain. *Human Brain Mapping*, 9(2):81–92, February 2000.

[95] Paul M. Thompson, Jay N. Giedd, Roger P. Woods, David MacDonald, Alan C. Evans, and Arthur W. Toga. Growth patterns in the developing brain detected by using continuum mechanical tensor maps. *Letters to Nature*, pages 190–193, March 2000.

[96] Paul M. Thompson, Michael S. Mega, Christine Vidal, Judith L. Rapoport, and Arthur W. Toga. Detecting disease-specific patterns of brain structure using cortical pattern matching and a population-based probabilistic brain atlas. In *Information Processing in Medical Imaging (IPMI)*, pages 488–501. Springer, June 2001.

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

[98] Paul M. Thompson and Arthur W. Toga. A framework for computational anatomy. *Computing and Visualization in Science*, 5:13–34, 2002.

[99] A. W. Toga. *An Introduction to Brain Warping.* Academic Press, San Diego, 1999.

[100] Arthur W. Toga and Paul M. Thompson. Temporal dynamics of brain anatomy. *Annual Review of Biomedical Engineering*, 5:119–145, 2003.

[101] Koen van Leemput, Frederik Maes, Dirk Vandermeulen, and Paul Suetens. Automated model-based bias field correction of MR images of the brain. *IEEE Transactions on Medical Imaging (TMI)*, 18(10):885–896, October 1999.

[102] Koen van Leemput, Frederik Maes, Dirk Vandermeulen, and Paul Suetens. Automated model-based tissue classification of mr images of the brain. *IEEE Transactions on Medical Imaging (TMI)*, 18(10):897–908, October 1999.

[103] Paul Viola and William Wells. Alignment by maximisation of mutual information. In *Proceedings of the 5th International Conference on Computer Vision (ICCV)*, pages 15–23, 1995.

[104] Simon K. Warfield, Kelly H. Zou, and William M. Wells. Validation of image segmentation and expert quality with an expectation-maximization algorithm. Lecture Notes in Computer Science (LNCS), pages 298–306. Springer, September 2002.

[105] W. Weinholt and B. Sendhoff. How to determine the redundancy of noisy chaotic time series. *International Journal of Bifurcation and Chaos*, 6(1):101–117, 1996.

[106] P P Wyatt and J A Nobel. Map mrf joint segmentation and registration of medical images. *Medical Image Analysis (MedIA)*, 7(4):539–552, 2003.

[107] Chen Xiaohua, Michael Brady, Jonathan Lo, and Niall Moore. Simulataneous segmentation and registration of contrast-enhanced breast mri. In *Information Processing in Medical Imaging (IPMI)*, number 3565 in Lecture Notes in Computer Science (LNCS), pages 126–137. Springer-Verlag, July 2005.

[108] Chen Xiaohua, Michael Brady, and Daniel Reuckert. Simulataneous segmentation and registration for medical image. In *Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 3216 of *Lecture Notes in Computer Science (LNCS)*, pages 663–670. Springer-Verlag, September 2004.

[109] A Yezzi and L Zollei. A variational framework for integrating segmentation and registration through active contours. *Medical Image Analysis (MedIA)*, 7:171–185, 2003.

[110] Jie Zhang and Anand Rangarajan. Affine image registration using a new information metric. In *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 848–855. IEEE, 2004.

[111] Jie Zhang and Anand Rangarajan. Multimodality image registration using an extensible information metric and high dimensional histogramming. In *Information Processing in Medical Imaging (IPMI)*, number 3565 in Lecture Notes in Computer Science (LNCS), pages 725–735. Springer-Verlag, July 2005.

[112] K. H. Zou, S. K. Warfield, J. R. Fielding, C. M. C. Tempany, W. M. Wells, M. R. Kaus, F. A. Jolesh, and R. Kikinis. Statistical validation based on parametric receiver operating characteristic analysis of continuous classification data. *Academic Radiology*, 10(12):1359–1368, December 2003.

[113] Karol Życzkowski. Renyi extrapolation of shannon entropy. *Open Systems and Information Dynamics*, 10:297, 2003.

Colophon

This dissertation was typeset using the LaTeX2e typesetting system with the UNC thesis style file (special thanks to John Keyser, Tom Hudson, and Michele Weigle). The document was composed in the LyX authoring system and the bibliography was created using BibDesk, a front end to BibTeX. Plots and figures were produced using Matlab and OmniGraffle.