Atlas-Based Hippocampus Segmentation In Alzheimer's Disease and Mild Cognitive Impairment

Owen T. Carmichael, Howard A. Aizenstein, Simon W. Davis, James T. Becker,
Paul M. Thompson, Carolyn Cidis Meltzer, Yanxi Liu

CMU-RI-TR-04-53

November 2004

Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Abstract

Purpose. To assess the performance of standard image registration techniques for automated MRI-based segmentation of the hippocampus in elderly subjects with Alzheimer's Disease (AD) and mild cognitive impairment (MCI). **Methods.** Structural MR images of 54 age- and gender-matched healthy elderly individuals, subjects with probable AD, and subjects with MCI were collected at the University of Pittsburgh Alzheimer's Disease Research Center. Hippocampi in subject images were automatically segmented by using AIR, SPM, FLIRT, and the fully-deformable method of Chen to align the images to the Harvard atlas, MNI atlas, and randomly-selected, manually-labeled subject images. Mixed-effects statistical models analyzed the effects of side of the brain, disease state, registration method, choice of atlas, and manual tracing protocol on the agreement between automated segmentations and expert manual segmentations. **Results.** Registration methods that produced higher degrees of geometric deformation produced automated segmentations with higher agreement with manual segmentations. Automated-manual agreement between Chen's method and expert manual segmentations were competitive with manual-manual agreement. Segmentations of the right hippocampus were more consistent with manual segmentations than those of the left. Automated-manual agreement was significantly lower in AD brains than MCI or controls. Automated segmentations based on registration with a randomly-selected subject image were more consistent with manual segmentations than those based on registration with the Harvard or MNI atlas. The manual tracing protocol was a significant source of variation in automated-manual agreement.

Contents

1 Introduction **1**

2 Related Work **4**

3 Methods **5**

3.1 Problem formulation 6

3.2 Atlas-based segmentation formulation 6

3.3 Registration methods 7

3.3.1 AIR 8

3.3.2 SPM 9

3.3.3 Chen's method 9

3.3.4 FLIRT 10

3.3.5 Summary 11

3.4 Registration algorithm details 11

4 Experiments **12**

4.1 Subject data 13

4.2 Manual segmentations 13

4.3 Cohort atlases 13

4.4 Standard atlases 14

4.5 Performance measures 14

4.5.1 Overlap ratio measures degree of agreement between segmentations 15

4.5.2 Closest-point distances measure severity of errors 16

4.5.3 Sectional analysis 16

4.6 Statistical analysis: mixed-effects models 17

5 Results **18**

5.1 Cohort atlases 18

5.2 Standard atlases and atlas masks 21

5.3 Cohort atlases vs. standard atlases 23

5.4 Comparing manual-automated agreement to manual-manual agreement 24

5.5 Sectional results 26

6 Discussion **28**

7 Conclusion **35**

1 Introduction

This paper presents a quantitative evaluation of methods for automatically delineating (or *segmenting*) the hippocampus in MR images of elderly subjects with Alzheimer's disease (AD) and mild cognitive impairment (MCI). We focus on *atlas-based* approaches that geometrically align (or *register*) the subject image to a reference image on which the hippocampus has been manually delineated. Our experiments evaluate the use of several widely-disseminated software registration packages for atlas-based hippocampus segmentation on images of 54 elderly controls, AD subjects, and MCI subjects.

Hippocampus segmentation in MCI and AD While the techniques we describe are general enough to apply to automated segmentation of arbitrary structures in the brain, we focus on the elderly hippocampus because it plays a critical role in the neurodegenerative progression of AD. In particular, hippocampal atrophy is known to occur early in the course of AD on a spatial scale large enough to be detectable with structural MR images [3] [33]. Therefore, hippocampal atrophy has been proposed as a clinical marker for early AD. Visual, qualitative atrophy assessment (see, for example, [15]) has been hindered by the presence of low-contrast boundaries between neighboring anatomical structures, varying protocols for atrophy assessment, and the relative subtlety of atrophy early in the course of AD [20]. However, the development of reliable, repeatable protocols for human raters to delineate the hippocampus have led to the possibility of precise quantitative evaluation of the hippocampus (e.g., [29] [27]). Besides enabling the quantitative study of atrophy in MCI and AD [7], hippocampus segmentation also enables region-of-interest-level quantification of hippocampus activation in functional images that have a co-registered structural image [16] and allows investigators to study other features of elderly hippocampi, such as their bilateral symmetry [2].

Automated methods However, manual segmentations are labor-intensive, vary from person to person, and require training the rater. Typical hippocampi take between 30 minutes and 2 hours to trace by hand; furthermore, expert raters quickly become fatigued by the manual dexterity and hand-eye coordination required to perform the segmentation, so it is usually not possible for them to spend long, continuous periods of time tracing. Each rater must be trained and validated by an expert, who must take care that hippocampus boundaries traced by the trainee are consistent with those of other raters. For these reasons, manual segmentation of large numbers of hippocampi for broad studies of atrophy effects has not been feasible. Several authors have proposed semi-automated segmentation methods that reduce manual segmentation labor by having the user identify a sparse set of image landmarks that constrain a subsequent automated segmentation process (see, for example, [46] [19] [8]). Here, however, we focus on fully-automated techniques to eliminate the need for a user to manually process each image under study, and to eliminate the landmark-identification process as a source of variability between segmentations of the same image. In so doing, we aim to overcome the difficulties in segmentation introduced by age- and AD-associated variability in hippocampal shape and volume.

Atlas-based segmentation We evaluate *atlas-based* techniques for automated segmentation of subject images. The key components are a special reference image called the *atlas image*, an *atlas mask*, *i.e.* a representation of the coordinates of the structure of interest in the atlas image, and a technique for image registration. Atlas-based segmentation simply registers the atlas image to the subject image, and uses the resulting spatial transformation to map the coordinates of the structure of interest from the atlas image to the subject image. Atlas-based segmentation, while conceptually simple, has several favorable qualities. First, the general problem of image registration is at the heart of a wide variety of medical applications including visualization, image-guided surgery, and voxel-based morphometry. This allows atlas-based segmentation techniques to take advantage of methodological advances driven by a wide range of application areas. Furthermore, atlas-based approaches are among the easiest to implement since they only require the user to align the atlas and subject images. Competing approaches to automatic hippocampus segmentation usually involve more complex optimizations involving prior models of how the the shape of the structure of interest varies over a population. Our view is that these more complicated approaches have the potential to provide slightly more accurate estimates of the structure of interest since they make use of more information than our atlas-based approach. Indeed, in the long term we envision an overall system in which atlas-based techniques provide an initial estimate of the subject mask, and then more complex procedures refine that estimate if needed for the application at hand.

Standard registration packages and transformation models We segment the hippocampus in a subject image by registering it to an atlas image on which the hippocampus has been traced manually. Previous studies have reported atlas-based hippocampus segmentation results based on recently-developed, cutting-edge registration algorithms that lack a widely-disseminated, standard software implementation (*e.g.*, [12]). Furthermore, no studies to date have systematically compared competing registration methods in terms of their performance in atlas-based elderly hippocampus segmentation. In contrast, our goal is to assess the accuracy of hippocampus segmentations that are generated using a variety of software packages that are already widely employed to process images at brain imaging laboratories. In particular, our experiments evaluate the registration components of the AIR [55], SPM [21], and FSL [28] packages, as well as an implementation of Chen's algorithm [6] that is similar to the classic Demons registration algorithm [48] (which is available in, *e.g.* the ITK software package [56]). We focus on established software packages to explore the possibility for investigators to immediately analyze brain structures in their rapidly-growing databases of geriatric MR images [1]. Besides comparing individual registration packages, our results also indicate general trends in terms of overall mathematical characteristics of their algorithms. In particular, our results suggest that atlas-based segmentation performance can

[1] Indeed, we have anecdotal evidence that many laboratories have already applied SPM and AIR to atlas-based elderly hippocampus segmentation, despite a lack of experimental validation

be greatly enhanced by registration methods that are allowed to geometrically deform the subject image to a high degree while registering it to the atlas image.

Standard atlases Our experiments also examine how the choice of atlas image and atlas mask affect atlas-based segmentation performance. In particular, we use the registration techniques to align the subject image to the Harvard and MNI atlases [31] [50]. We chose these atlases– our *standard atlases*– for the same reason we chose our registration algorithms– they are already widely used in brain imaging laboratories to define a common reference frame for population-based inferences in voxel-based morphometry [1], deformation-based morphometry [9], and related methods. As opposed to probablistic atlases [11], which are averaged representations of multiple brains, the Harvard and MNI atlases are derived from one or more images of a single subject. Standard atlases can be advantageous for atlas-based segmentation since they contain a well-studied image and an extensive set of manual segmentations, and therefore require no hand-labeling of images on the part of the user. However, the use of standard atlases may cause registration difficulties if the standard atlas image differs widely from the subject images. In particular, standard atlas images are usually scans of young, healthy brains; age-associated or disease-associated structural characteristics in the subject image can make the problem of registering it to the atlas image so difficult that registration can fail. Furthermore, differences in signal characteristics between the atlas image and subject images can add difficulty to the registration process. Another potential difficulty with standard atlases is that their atlas masks are based on a set of anatomical boundary conventions that may not coincide with the conventions of a particular laboratory.

Cohort atlases To address these difficulties, we consider an alternative segmentation approach in which the user manually segments the structure of interest on one or more subject images selected from a population. The subject images, augmented with the manual segmentations, are then treated as ”atlas images;” that is, they are used as the reference images that all other images in the population are registered to in order to automatically segment the structure of interest in the rest of the population. We refer to the selected, manually-segmented subject images as *cohort atlas images* since they are drawn from the same cohort as the other subject images to be segmented. While it is unusual to refer to individual subject images with manual segmentations as ”atlases,” we use the term to emphasize their role in providing a standard reference frame for aligning images of interest to during automated segmentation. While cohort atlas images reflect characteristics that are peculiar to a particular scan of a particular subject, we feel that they have potential advantages over standard atlases. If the population of images is homogeneous with respect to factors such as sensor acquisition parameters, subject age, and subject disease state, then drawing a cohort atlas image from the population guarantees that these factors will not confound the registration process. Furthermore, hand-labeling the structure of interest insures the user that anatomical boundaries reflect his or her conventions. A drawback of cohort atlases is that they are inherently more labor-intensive than standard atlases since they require the user to hand-label the structure of interest on each selected cohort atlas image. Standard atlases and cohort atlases have both been employed by a variety of atlas-based segmentation techniques (see, *e.g.*, [26] *vs.* [54]).

The rest of this paper is organized as follows. We give a broad overview of previous approaches to hippocampus segmentation, and atlas-based segmentation in general, in Section 2. Section 3 describes the registration methods we evaluate in terms of a unified mathematical formulation. In Section 4 we describe the set of images our algorithms are applied to, the manual segmentations we use for validation, and the design of our experiments. Section 5 presents results of the experiments, followed by a discussion in Section 6.

2 Related Work

In this section we differentiate our work from prior studies in atlas-based segmentation and hippocampus segmentation in general. Our emphasis on quantitative evaluation of a difficult atrophy-affected brain structure, and comparison of several standard methods, sets our study apart from related work.

Quantitative validation of elderly hippocampus segmentation Our study augments the limited number of quantitative evaluations of fully-automated hippocampus segmentation algorithms applied to AD subjects. Fischl *et al.* [18] applied their automated brain segmentation technique to subject groups that roughly correspond to our control, MCI, and AD groups, and showed that differences in automatically-estimated hippocampal volumes between groups correspond well with expected AD-related atrophy. Similarly, Freeborough *et al.* [19] indirectly validate their semi-automated segmentations by showing that estimated hippocampal volume decreases rapidly in serial scans of autosomal AD subjects. Here, we go a step further by directly validating the quality of automatically-segmented AD, MCI, and control hippocampi against manual segmentations by expert raters. Crum *et al.* [12] segmented the hippocampus in a subject image by registering it to a manually-segmented image of the same subject taken roughly a year previously, and validate the approach on serial scans of elderly control and AD subjects. Our experiments with cohort atlases take a similar approach to validating inter-subject, as opposed to intra-subject, segmentation. Rizzo *et al.* [43] quantitatively validated their standard-atlas-based segmentation technique on a single AD subject, along with a Parkinson's disease subject and 3 controls. Perez de Alejo *et al.* [14] evaluated a semi-automated method on a limited number of hippocampus slices in an AD population. In terms of quantitative evaluations of elderly hippocampus segmentation in general, Shen *et al.* [46] have shown that their semi-automated technique is competitive with manual tracing on the hippocampi of a set of 10 elderly subjects. Several authors, including Haller *et al.* [23], have validated their segmentation approaches by checking the agreement between the volumes of automated and manual segmentations.

Hippocampus segmentation in general Other automated approaches have been applied to segmenting the hippocampus in younger subjects, especially those with schizophrenia and other hippocampus-relevent conditions. While conditions such as schizophrenia can introduce variability in hippocampal shape and volume that complicate automated segmentation, young adult hippocampi are generally larger, rounder, and more distinct from surrounding tissue than elderly hippocampi. Webb *et al.* [54] applied a semi-automated, atlas-based segmentation technique to subjects with temporal lobe epilepsy (TLE), showing that the resulting hippocampal volume can be a marker for TLE-related atrophy. Several recent automated methods applied to young hippocampi have combined statistical models of its appearance with prior models of its expected shape; these techniques vary mainly in terms of whether shape variation is represented in terms of surfaces [30] [40], medial structures [41], or dense volume deformations [17]. Our view is that these more complex shape-and-appearance based methods have the potential in the future to be extended so that they can accurately deal with the complications of age-associated and AD-associated atrophy. Indeed, we envision using atlas-based techniques to provide an initial starting guess at the segmentation, which is further refined by a more precise shape and appearance model.

Since fully-automated hippocampus segmentation is relatively difficult, semi-automated methods are more common. In low-level semi-automated techniques, the user indicates starting points or constraints for low-level image processing routines like region-growing or pixel clustering [19] [4]; in contrast, high-level techniques allow the user to provide constraints or initial conditions for alignment between the subject image and an atlas image or shape model [46] [25] [8] [23]. In other semi-automated techniques, the segmentation of the hippocampus on an initial slice is propagated to other slices in the volume [14]. Here, we focus on the long-term goal of fully-automated hippocampus segmentation to eliminate required interaction with a human user.

Atlas-based segmentation in general Atlas-based segmentation techniques can be used to segment any structure or tissue type of interest that has been manually segmented on the atlas image. Therefore, while a variety of authors have applied atlas-based segmentation to other brain structures and subject groups (for example, [51] [10] [5] [34]) it is possible that their methods could be applied to the hippocampus in elderly dementia patients. However, other atlas-based segmentation techniques are generally based on institution-specific implementations of recent algorithms, while we focus on packages such as SPM and AIR that are already widely disseminated and validated. Dawant and colleagues validated an atlas-based segmentation technique similar to Chen's method, which we evaluate here, on a limited number of cerebellum slices on a set of severely atrophied alcoholic brains [13] [24]. Our analysis of Chen's method extends these experiments to entire atrophied hippocampi.

3 Methods

In this section we provide a mathematical framework for atlas-based segmentation through which we describe the standard algorithms we compare experimentally. While several operating characteristics vary between each algorithm, we highlight the degree to which each method is allowed to geometrically deform the subject image while matching it to the atlas. This degree of distortion is summarized in the *geometric transformation model*.

3.1 Problem formulation

We assume we are given a 3-dimensional *subject image*, **I**, where $\mathbf{I}(x, y, z)$ represents the image intensity at voxel location $[x, y, z]$ [2]. Our goal is to process **I** and recover a binary volume **B**, where $\mathbf{B}(x, y, z) = 1$ if **I**(x,y,z) is located in the hippocampus of the brain, and $\mathbf{B}(x, y, z) = 0$ otherwise. We refer to **B** as the *structure mask* for **I**. We focus on this representations of the hippocampus for its simplicity and usefulness in volumetric analysis; however we note that other representations, such as parametric surfaces [30], may be more appropriate target outputs for some applications.

3.2 Atlas-based segmentation formulation

We assume that we have access to one or more atlas images $\mathbf{I}_t$. In this paper we only consider the use of a single atlas image, however we note that other studies have considered the use of multiple atlases (*e.g.*, [44]). Along with the atlas image, we assume we have access to an atlas mask $\mathbf{B}_t$, where $\mathbf{B}_t(x, y, z) = 1$ if $\mathbf{I}_t(x, y, z)$ is located in the structure of interest of $\mathbf{I}_t$, and $\mathbf{B}_t(x, y, z) = 0$ otherwise. Atlas-based segmentation proceeds by first estimating the parameters of a geometric transformation between the subject and atlas images, and then using the estimated transformation to relate the atlas structure mask to the subject structure mask. In more detail, we assume that an atlas image $\mathbf{I}_t$ and subject image **I** are related to each other by the following model:

$$\mathbf{I}(g_\phi(x, y, z)) = h_\psi(\mathbf{I}_t(x, y, z)) + \gamma$$

The geometric transformation model g_ϕ transforms voxel locations in the atlas image to voxel locations in the subject image; its behavior is governed by a vector of parameters, ϕ. For example, if we assume a rigid geometric transformation model, ϕ will have entries for rotation angles about, and translations along, each of the three cardinal axes. The intensity transformation model h_ψ relates the intensities of corresponding voxels in the atlas and subject images; it is meant to account for signal characteristics, such as gain and field inhomogeneities, that differ between images. An example intensity transformation is a linear scaling of intensity values, for which the parameter vector ψ governing the behavior of the intensity transformation has only one entry, namely the scaling parameter. In atlas-based segmentation, we first use optimization techniques to estimate ϕ and ψ, and then return an estimate $\hat{\mathbf{B}}$ of **B** by applying g_ϕ to $\mathbf{B}_t$:

$$\hat{\mathbf{B}}(x, y, z) = \mathbf{B}_t(g_\phi(x, y, z))$$

A schematic illustration of atlas-based segmentation is shown Figure 1.

[2] Matrices are written in uppercase bold, and their elements are indexed in parentheses, *e.g.* voxels in image volumes (*i.e.* 3D matrices) are denoted $\mathbf{A}(x, y, z)$. 1-D matrices (*i.e.*, vectors) with entries $x, y, z, ...$ are written $[x, y, z, ...]$. Scalar-matrix multiplication is denoted with $\cdot$: $a \cdot [x, y, z] = [ax, ay, az]$. Matrix-matrix multiplication is written by juxtaposing the two matrices.

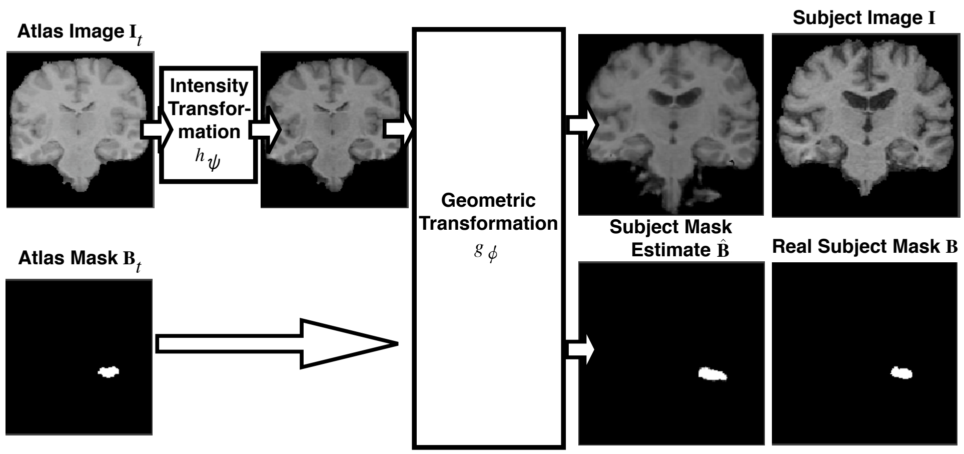
Atlas Image $\mathbf{I}_t$
Intensity Transformation
h_ψ
Geometric Transformation
g_ϕ
Subject Image $\mathbf{I}$
Subject Mask Estimate $\hat{\mathbf{B}}$
Real Subject Mask $\mathbf{B}$
Atlas Mask $\mathbf{B}_t$

Figure 1: Schematic view of atlas-based segmentation. An intensity transformation and geometric transformation are estimated to register the atlas image to the subject image; the geometric transformation is applied to the atlas mask in order to estimate the subject mask.

3.3 Registration methods

We compare the performance of AIR, SPM, FLIRT, and Chen's method as registration substrates for atlas-based segmentation of elderly hippocampi. Here, we briefly describe each of these methods in terms of their most important components, which are:

- **Geometric Transformation Model:** This is the functional form chosen for g_ϕ.
- **Intensity Transformation Model:** This is the functional form chosen for h_ψ.
- **Cost function:** The cost function gives a numerical score to putative solutions for ϕ and ψ. Given a particular solution for ϕ and ψ, it can be thought of as a function $c(\phi, \psi)$ that in some way compares the image intensities $\mathbf{I}(g_\phi(x, y, z))$ and $h_\psi(\mathbf{I}_t(x, y, z)) + \gamma$ for some number of image locations (x, y, z), and returns lower values when the intensities are more similar to each other.
- **Optimizer:** Given choices for g_ϕ, h_ψ, and $c(\phi, \psi)$, the optimizer is the numerical technique used for finding solutions for ϕ and ψ that minimize $c(\phi, \psi)$.
- **Multi-scale strategy:** Most methods incorporate schemes for smoothing and/or downsampling $\mathbf{I}_t$ and $\mathbf{I}$ before evaluating $c(\phi, \psi)$. Doing so serves to smooth $c(\phi, \psi)$, and also to help ease the computational burden of evaluating it.

In summary, each of the registration techniques may be summarized at a high level as employing an optimizer to adjust the parameters of geometric and intensity transformations between atlas and subject images so that doing so minimizes a cost function. For AIR, SPM, and Chen's method, this optimization is performed in a series of stages, where each subsequent stage corresponds to a more complex geometric transformation

	Geometric	**Intensity**	**Optimizer**
AIR affine	Affine	Linear scaling	Newton
SPM affine	Affine	Linear Scaling	Newton
FLIRT affine	Affine	None	Powell's Method
AIR semi-deformable	Polynomial basis functions	Linear scaling	Newton
SPM semi-deformable	DCT basis functions	Linear scaling	Newton
Chen semi-deformable	Piecewise linear	Mean and variance normalization	Levenberg-Marquardt
Chen fully-deformable	Dense voxel flow	Mean and variance normalization	Gradient descent

Table 1: Algorithmic properties of the compared registration methods.

model. Table 1 briefly summarizes the geometric and intensity transformations, optimizer, and multi-scale strategy for each stage of the methods we compare in our experiments. Each method we evaluate contains additional algorithmic details that are not captured by these components; however, we feel that the components capture the most important aspects in which the overall methods operate and differ from each other. Whenever possible, we set other algorithmic parameters to be identical from package to package. For example, for all packages we used a trilinear model to interpolate image intensities to sub-voxel locations, and we chose a sum-of-squared-differences (SSD) cost function. We chose SSD because the images in our population were all acquired on the same scanner with similar imaging parameters, meaning that the global distribution of intensities does not vary significantly from image to image. For this reason, we did not expect that cost functions designed to capture complex relationships between image intensities from image to image (mutual information, for example [42]) would significantly improve our registration results. We note that while all methods employed SSD, FLIRT used an apodized version; that is, voxels closer to the edge of the overlapping brain region were weighted lower than those closer to the center.

3.3.1 AIR

The first stage of AIR estimates the parameters of an affine geometric transformation. That is, $g_\phi(x, y, z) = \mathbf{A}[x, y, z, 1]$ for a 4-by-4 matrix **A** determined by 12 independent parameters. The intensity transformation model is a linear scaling, $h_\psi(\mathbf{I}_t(x, y, z)) = w \cdot \mathbf{I}_t(x, y, z)$, with a single parameter w. This first stage of AIR estimates the 12 parameters of **A** along with w by a Newton-type iterative optimizer. In the second stage of AIR, the geometric transformation model consists of projecting the spatial coordinates onto a polynomial basis of degree K, specifically:

$$g_\phi(x, y, z) = \sum_{p=0}^{K} \sum_{q=0}^{K} \sum_{r=0}^{K} [a_{pqr1}, a_{pqr2}, a_{pqr3}] \cdot x^p y^q z^r$$

The coefficients a_{pqr} are the geometric transformation parameters; they are estimated by the same Newton-type minimization as in the affine case. The degree K of the polynomial basis is a user-set parameter; however, AIR allows the user to estimate the transformation parameters successively for increasing values of K, using the solution for $K = k - 1$ as the starting point for estimating parameters for $K = k$. Using this setting, we increase K from 2 to 12, terminating the estimation early if the Newton minimization becomes ill-conditioned. In both stages of AIR, the multi-scale strategy is to compute the cost function at every k-by-k-by-kth voxel, where k is increased by factors of 3 over the course of optimization, from 81 to 1.

3.3.2 SPM

As in AIR, the first stage of SPM estimates a 12-parameter affine geometric transformation and a single scaling parameter for the intensity transformation. For the second stage, the geometric transformation model follows the same functional form as a discrete cosine transform (DCT), that is,

$$g_\phi(x,y,z) = [x,y,z] + \sum_{p=0}^{K} [a_{p1}, a_{p2}, a_{p3}] \cdot d_p(x,y,z)$$

The functions $d_p(x,y,z)$ are the low-dimensional basis functions of the DCT, and the problem is to estimate the coefficients a_{p1}, a_{p2}, a_{p3}. In both stages, the parameters are estimated using a Gauss-Newton minimization procedure. The multiscale strategy employed by SPM is to evaluate the cost function at every k-by-k-by-kth voxel as in AIR; however, rather than setting a fixed, prior schedule for k, SPM modulates k at each iteration of the optimization procedure according to the error in voxel intensities between the aligned atlas and subject images at that iteration. Specifically, SPM computes the variance in $(\mathbf{I}(g_\phi(x,y,z)) - h_\psi(\mathbf{I}_t(x,y,z)))^2$ at each iteration, and sets k proportional to that variance. In so doing, SPM samples more finely as iterations proceed and the intensity error variance reduces.

With respect to our comparison of registration techniques, it is important to note that SPM explicitly biases its geometric transformation parameter estimates toward transformations that deform the subject volume minimally. Specifically, SPM simultaneously attempts to minimize the SSD error between the geometrically- and intensity-aligned subject and atlas images, as well as the magnitude of the DCT coefficients a_{p1}, a_{p2}, a_{p3}. The bias toward minimally-deforming transformations is motivated by the application of SPM to the spatial normalization of images for voxel-based morphometry (VBM). However, for our atlas-based segmentation application, the goal is to deform the subject image so that the voxels of the subject hippocampus aligns as well as possible with the voxels of the atlas hippocampus, regardless of how heavily the subject image needs to be deformed; thus, SPM may be at a fundamental disadvantage against AIR and Chen's method, since they both encourage high-quality image alignment with no bias toward minimally-deforming transformations. However, we include SPM in our results because various authors routinely employ SPM for atlas-based segmentation purposes.

3.3.3 Chen's method

The registration method of Chen consists of three stages that estimate similarity, piecewise-linear, and dense voxel-by-voxel geometric transformations respectively. In the first stage, a translation, rotation, and scaling between the images is estimated, and is used as the starting point for the estimation of a piecewise-linear geometric transformation.

The piecewise-linear model is specified in terms of the 3d coordinates of a set of control points $\{[x_g, y_g, z_g]\}$, and displacements of the control points, $\{[\delta x_g, \delta y_g, \delta z_g]\}$. The control points form a regular 3D rectangular grid that covers the atlas image, so that each voxel $[x, y, z]$ in the atlas image can be categorized as belonging to a sub-volume, or *cell*, bounded by eight control points : $[x_{gl}, y_{gl}, z_{gl}]$, $[x_{gh}, y_{gl}, z_{gl}]$, $[x_{gl}, y_{gh}, z_{gl}]$, $[x_{gl}, y_{gl}, z_{gh}]$, $[x_{gh}, y_{gh}, z_{gl}]$, $[x_{gh}, y_{gl}, z_{gh}]$, $[x_{gl}, y_{gh}, z_{gh}]$, $[x_{gh}, y_{gh}, z_{gh}]$, such that $x_{gl} < x < x_{gh}$, $y_{gl} < y < y_{gh}$, $z_{gl} < z < z_{gh}$. The geometric transformation for each voxel is a trilinear interpolation of the displacements of the control points the bound its cell. That is,

$$g_\phi(x, y, z) = [x, y, z] + [\alpha_x * \delta x_{gl} + (1 - \alpha_x) * \delta x_{gh}, \alpha_y * \delta y_{gl} + (1 - \alpha_y) * \delta y_{gh}, \alpha_z * \delta z_{gl} + (1 - \alpha_z) * \delta z_{gh}]$$

where $\alpha_x = (x_{gh} - x)/(x_{gh} - x_{gl})$ and similarly for α_y and α_z. The parameters to estimate for this geometric transformation are the control point displacements $[\delta x_g, \delta y_g, \delta z_g]$. Chen's method uses the Levenburg-Marquardt method to iteratively estimate these parameters. Instead of estimating a single piecewise linear transformation, Chen's method estimates a series of piecewise linear transformations corresponding to increasingly fine-grained grids of control points. In other words, the method first estimates displacements for a 2x2x2 grid of control points, uses these displacements as the starting point for computation of displacements of a 3x3x3 grid, and so on. The piecewise-linear geometric transformation is the starting point for the estimation of a dense voxel-by-voxel transformation in the third stage of Chen's method, or in other words,

$$g_\phi(x, y, z) = [x, y, z] + [\delta x, \delta y, \delta z]$$

The parameters to estimate for this transformation model are the 3D displacements $[\delta x, \delta y, \delta z]$ for each voxel $[x, y, z]$ in the atlas image. The displacements are estimated using a method similar to the Demons algorithm of Thirion *et al.*. Specifically, a first-order Taylor expansion of the constraint $\mathbf{I}((x + \delta x, y + \delta y, z + \delta z)) = h_\psi(\mathbf{I}_t(x, y, z))$ yields an equation that gives an appropriate displacement $[\delta x, \delta y, \delta z]$ solely in terms of the images $\mathbf{I}_t$ and $h_\psi(\mathbf{I}_t)$, and their spatial image gradients. This displacement is computed iteratively at each voxel in the atlas image until it converges.

Chen's intensity transformation model consists of a translation and scaling, *i.e.* $h_\psi(x) = ax + b$. The parameters, a and b, are estimated separately from the geometric transformation parameters, using the simple heuristic that the mean and variance of the intensity distribution of the atlas image should match that of the subject image. For the piecewise-linear stage, the cost function is computed at a random selection of k voxels; while in principle we could increase k as the granularity of the grid of control points becomes finer, we simply keep a constant k throughout the piecewise linear stage.

3.3.4 FLIRT

Like the first phases of AIR and SPM, FLIRT estimates an affine geometric transformation using a multiscale strategy that computes a cost function at every k-by-k-by-k voxels, with k increasing over the course of the optimization. However, FLIRT uses a significantly different overall optimization strategy, cost function, and intensity transformation model. FLIRT apodizes its cost function; that is, at each iteration of estimating ϕ, the cost function depends on the overlap between the brain portions of $g_\phi(\mathbf{I})$

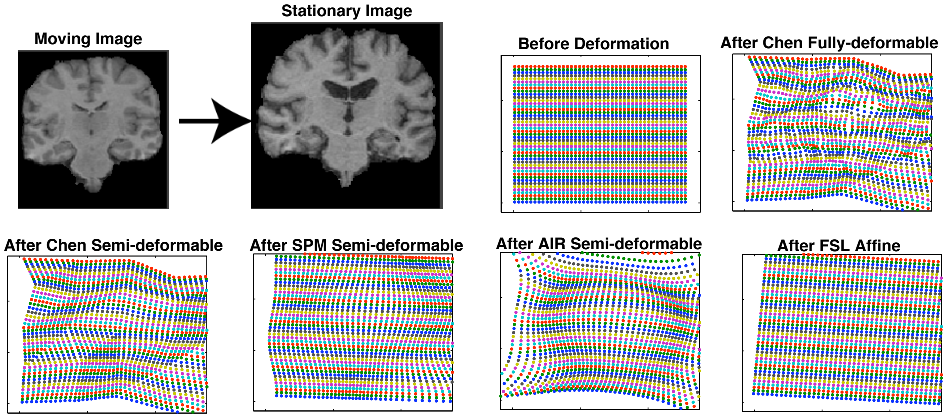
Moving Image
Stationary Image
Before Deformation
After Chen Fully-deformable
After Chen Semi-deformable
After SPM Semi-deformable
After AIR Semi-deformable
After FSL Affine

Figure 2: Example image deformations produced by fully-deformable, semi-deformable, and affine registration techniques. The moving image is registered to the stationary image using each of the 7 algorithms we analyze. The colored dots show the geometric positions of voxels in the shown slice of the moving image before and after deformation by each of the methods. The transformation produced by the AIR affine method and SPM affine method were almost identical to that of the FSL affine method.

and $h_\psi(\mathbf{I}_t)$. Apodization amounts to downweighting the contribution that voxels near the edge of the overlapping region make toward the overall cost function. Furthermore, FLIRT simultaneously maintains multiple estimates of ϕ, each derived from its own random initial guess; over the course of the optimization, these competing estimates are winnowed down to a single, final answer for ϕ. Unlike AIR and SPM, FLIRT finds low-cost settings for ϕ using Powell's method, which is a downhill-simplex type approach that computes no derivatives. The final difference between FLIRT and the other afffine methods is that it incorporates no intensity transformation model.

3.3.5 Summary

The chief characteristics of these algorithms are summarized in Table 1. We refer to the basis function phases of SPM and AIR, as well as the piecewise-linear phase of Chen's method, as ”semi-deformable” methods, because while their geometric transformations vary spatially, they do so in a gradual, constrained, low-dimensional way; on the other hand, we refer to the dense voxel-by-voxel phase of Chen's method as a ”fully-deformable” method because the geometric transformation is fully unconstrained. An example showing the use of these techniques to register a pair of images in our MCI data set is shown in Figure 2. Note that the geometric transformation produced by the semi-deformable techniques is more spatially smooth than that of the fully-deformable technique.

3.4 Registration algorithm details

Our chief goal is to examine the impact of the factors listed in Table 1 on the quality of atlas-based hippocampus segmentation. Therefore, we strove to equalize all other operating parameters across all registration packages as much as possible. We used the sum-of-squared-differences (SSD) cost function to evaluate the quality of image alignment. Since all the images in our study were acquired with the same imaging modality and have similar intensity characteristics, we did not feel that more general, more computationally complex cost functions such as mutual information would significantly improve registration results. We also used a trilinear model to interpolate image intensities at sub-voxel locations. Each software package employed a slightly different criterion to determine when the iterative search for good values for ϕ and ψ should halt; whenever possible (AIR, SPM, Chen), we set the maximum number of iterations to 50. Each package also employed slightly different strategies for down-sampling the images prior to computing the cost function; for AIR, SPM, and FLIRT, the cost function is computed at every k-by-k-by-kth voxel, where k decreases over the course of optimization. For Chen's method, the cost function is computed at a random sample of k voxels at each iteration. Empirically, we found that near the beginning of the numerical optimization, the techniques varied widely in terms of how many voxel values were used to compute the cost function; however, near the end of optimization, all methods computed the cost function at a number of voxels corresponding to every k-by-k-by-kth voxel, where k varied between roughly 2 and 4.

4 Experiments

Our experiments evaluate the degree to which segmentation results vary with respect to disease state, registration algorithm, atlases, manual tracings, and side of the brain. At the core of our experiments is the following sequence of actions:

1. Registering an atlas image to a subject image
2. Using the resulting geometric transformation to transfer manually-labeled left and right hippocampus masks from the atlas image to the subject image
3. Evaluating the consistency between the resulting subject mask estimates and ground-truth manual tracings

We refer to the execution of these actions for a particular choice of atlas image, subject image, registration algorithm, and manual tracings as a segmentation *trial*. Our experimental results were obtained by performing a series of trials through which each of these 4 factors is varied systematically. In particular, for both of our standard atlases, we ran one trial for each possible combination of the 7 registration algorithms, 54 subject images, and 2 sets of manual tracings supplied with the atlas. Section 4.2 describes our acquisition of ground-truth manual tracings and explains why each atlas image is equipped with two distinct manual tracings of the left and right hippocampus. For the cohort atlas scenario, we group the images by disease state (AD, MCI, or control). For each disease state, and for each registration algorithm, we run one trial for each possible cohort atlas image and subject image within the disease group. The choice of cohort atlas image is described in Section 4.3, and a description of our subject images for the AD, MCI, and control populations is in Section 4.1. Section 4.5 describes the numerical measures we use to quantify the agreement between an estimated subject mask and the corresponding ground-truth mask. Results from the trials are statistically analyzed to determine the significance of the following factors on the consistency measures: registration algorithm, disease state, atlas type (standard or cohort), side of the brain, and choice of manual tracing on the atlas.

4.1 Subject data

Our subject data consists of MR images of 20,19, and 15 subjects in the AD, MCI, and control populations respectively. All subjects were enrolled in the University of Pittsburgh Alzheimer's Disease Research Center between 1999 and 2004 and given a structural MR scan at time of enrollment. The spoiled gradient-recalled (SPGR) volumetric T1-weighted pulse sequence, acquired in the coronal plane, has the following parameters optimized for maximal contrast among gray matter, white matter, and CSF (TE=5, TR =25, flip angle = 40 degrees, NEX = 1, slice thickness = 1.5mm/0mm interslice). Along with the MR scan, subjects received a comprehensive battery of neuropsychological and clinical tests at time of enrollment and at yearly follow-up visits (see [35] [36] for evaluation procedure). A consensus meeting of neuroradiologists, psychiatrists, neurologists, and psychologists diagnosed each subject into MCI [38], AD, or control categories.

Skulls were stripped from all images using the Brain Extraction Tool (`BET`) [47], and the images were cropped to remove all-zero slices using the `crop` tool provided with `AIR 2.0` [55].

4.2 Manual segmentations

We evaluate automated segmentations by comparing them to manual segmentations performed by a single expert rater, R1, who was blind to diagnosis, gender, age, and other clinical data at the time of tracing. Hippocampi were traced on contiguous coronal slices following the guidelines of Watson *et al.* [53], Schuff *et al.* [45], and Pantel *et al.* [37]. The traced structure included the hippocampus proper, the subiculum, and the dentate gyrus. The image and tracing were viewed in all three orthogonal viewing planes during manual segmentation. Addtionally, we selected 2 AD, 2 MCI, and 2 control images from the pool of 54 subjects for tracing by two additional trained raters, R2 and R3, using the same protocol. These additional manual segmentations were used to compare automated segmentation performance to inter-rater agreement. All manual segmentations were digitized into binary volumes for analysis.

4.3 Cohort atlases

In the cohort atlas scenario, we select an image– the cohort atlas image– from a subject population (AD, MCI, or control), manually trace left and right hippocampi on it, and automatically segment the hippocampi in all other images in that population by registering them to the cohort atlas image. An immediate question is how to select a cohort atlas image from the population. It may be possible to browse the entire collection of images and select one or more subject images that possess characteristics that are typical for the population; or, if the population is especially large, the user might simply select the cohort atlas image at random. Exploring the question of how to select an atlas image that is typical of a population, or in some way favorable for atlas-based segmentation, is beyond the scope of this paper (however, see [44] for an investigation of this issue). Therefore, we consider random selection of cohort atlases. In particular, for each image in each subject population, we consider a hypothetical situation in which that image is selected as the cohort atlas; all other images in the population are registered to the cohort atlas image and hippocampus segmentation results are evaluated. In other words, for a population of k images, we consider k different possible cohort atlases, which we register to all $k-1$ other images in the population for a total of $k*(k-1)$ trials per registration method.

4.4 Standard atlases

In the standard atlas scenario, we are given an atlas image and atlas masks provided by the atlas institution (Harvard or MNI). We register a subject image to the atlas image to segment its hippocampi, and evaluate the segmentation by comparing it to the manual segmentation performed by rater R1. However, we recognize that the manual segmentation protocol used by R1 may differ from that used by manual tracers at MNI and Harvard, and that our evaluation risks confounding two distinct sources of error: the automated algorithm and discrepancies between tracing protocols. For this reason, rater R1 traced left and right hippocampi on the Harvard and MNI atlas images, and automatically segmented subject hippocampi by transforming the R1-traced structures to the subject image.

4.5 Performance measures

We register the subject image to the atlas image in order to arrive at an estimate $\hat{\mathbf{B}}$ of the underlying hippocampus mask **B**. For any hippocampus mask **B**, we refer to the voxels in **B** that correspond to a portion of the hippocampus (*i.e.*, (x, y, z) such that $\mathbf{B}(x, y, z) = 1$) as the *structure voxels* of **B**. We wish to evaluate the agreement between $\hat{\mathbf{B}}$ and **B** by answering two questions: first, to what degree do the hippocampi in $\hat{\mathbf{B}}$ and **B** overlap with each other? Second, for the portions of the hippocampi in $\hat{\mathbf{B}}$ and **B** that are in error– *i.e.*, that do not overlap with each other– how far are they from overlapping? The first question aims to count the sheer number of voxels in $\hat{\mathbf{B}}$ and **B** that disagree with each other; the second question delves deeper into how extreme the errors are. Section 4.5.1 describes the overlap ratio, our criterion for quantifying the number of error voxels bewteen the two masks; Section 4.5.2 describes our use of closest-point distances (CPDs) to quantify the distance of error voxels to the correct hippocampal surface. When evaluating automated hippocampus segmentations, we feel it is important to quantify both the number of voxels between $\hat{\mathbf{B}}$ and **B** that are in disagreement, and how far the hippocampus voxels in $\hat{\mathbf{B}}$ are from the hippocampus surface in **B**. Both measures may be important to consider when evaluating atlas-based segmentation for applications in which a degree of hippocampus localzation error may be tolerable.

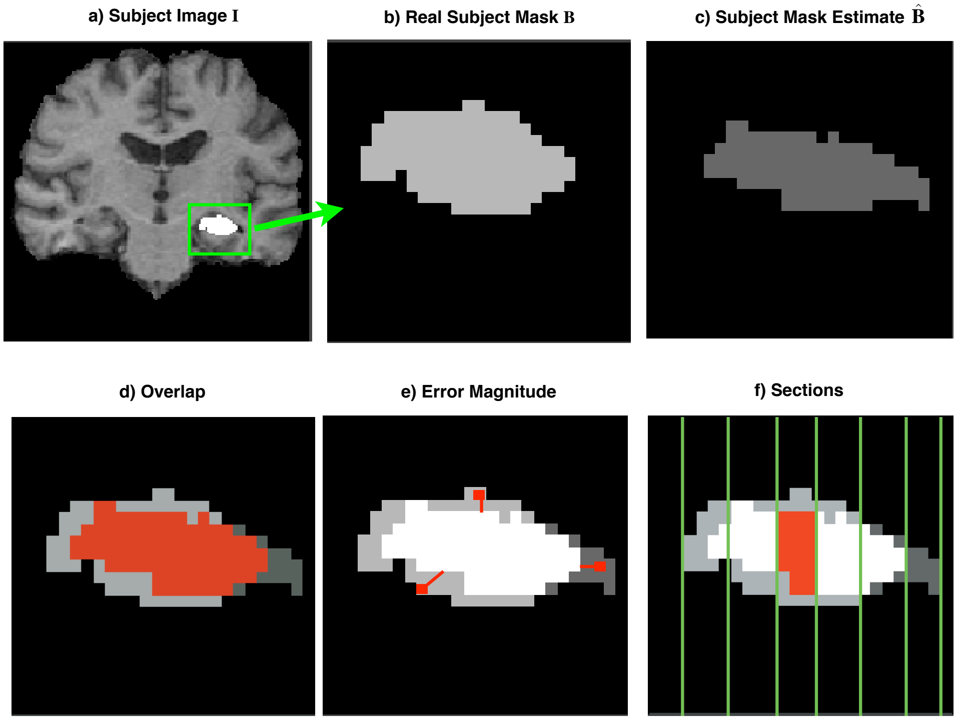
a) Subject Image $\mathbf{I}$
b) Real Subject Mask $\mathbf{B}$
c) Subject Mask Estimate $\hat{\mathbf{B}}$
d) Overlap
e) Error Magnitude
f) Sections

Figure 3: Evaluating consistency between masks using overall and sectional overlap and closest-point distances. A ground-truth subject mask and estimated subject mask are shown in light and dark gray. Figure 3d) : Voxels in red overlap between the ground-truth and the estimate. Overlap ratio measures the ratio between the volume of the red region and the volume of the combined red and gray regions. Figure 3e) : For each error voxel (in gray), the closest point distance measures the distance between the voxel and the surface of the other mask. Figure 3f) : The green bars split the hippocampus voxels into axis-parallel sections. In sectional analysis, overlap ratio and closest-point distances are computed for each section independently.

4.5.1 Overlap ratio measures degree of agreement between segmentations

To compute the overlap ratio, we consider three different sets of voxels: set $\mathcal{A}$ is the voxels that are labeled as hippocampus by both $\hat{\mathbf{B}}$ and $\mathbf{B}$; set $\mathcal{B}$ has voxels labeled as hippocampus by $\hat{\mathbf{B}}$ but not $\mathbf{B}$; and set $\mathcal{C}$ consists of voxels labeled as hippocampus by $\mathbf{B}$ but not $\hat{\mathbf{B}}$ (Sets $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ are labeled in white, dark gray, and light gray in Figure 3e). The *overlap ratio* for the two masks is computed as follows:

$$or(\mathbf{B}, \hat{\mathbf{B}}) = \frac{|\mathcal{A}|}{|\mathcal{A}| + |\mathcal{B}| + |\mathcal{C}|}$$

In other words, the overlap ratio measures the volume of ”$\mathbf{B}$ AND $\hat{\mathbf{B}}$” divided by the volume of ”$\mathbf{B}$ OR $\hat{\mathbf{B}}$”. When the two masks overlap perfectly, $or(\mathbf{B}, \hat{\mathbf{B}}) = 1$ since $\mathcal{B}$ and $\mathcal{C}$ are both empty; when the masks do not overlap at all, $or(\mathbf{B}, \hat{\mathbf{B}}) = 0$ since $\mathcal{A}$ is empty. The overlap ratio gives an easily interpretable measure of the degree to which the masks overlap; it gives the percentage of hippocampus voxels from the two masks that agree with each other. We note that several authors have quantified their automatic structure segmentation results in terms of similar criteria based on the relative sizes of sets $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$; see, for example, [13] [30] [46] [32].

4.5.2 Closest-point distances measure severity of errors

The overlap ratio gives a measure of the sheer number of hippocampus voxels in $\hat{\mathbf{B}}$ and **B** that agree with each other; for a more detailed picture of how the discrepancies between $\hat{\mathbf{B}}$ and **B** are distributed, we compute closest-point distances (CPDs) between error voxels and the hippocampus surfaces they should coincide with (see Figure 3e). That is, for each voxel $[x, y, z]$ in the set $\mathcal{B}$ above, we compute the distance between $[x, y, z]$ and the closest hippocampus voxel in **B**:

$$cp([x,y,z],\mathbf{B}) = \min_{\mathbf{B}(x_\mathbf{B},y_\mathbf{B},z_\mathbf{B})>0} d([x,y,z],[x_\mathbf{B},y_\mathbf{B},z_\mathbf{B}])$$

Similarly, for each voxel $[x, y, z]$ in set $\mathcal{C}$ above, we compute $cp([x, y, z], \hat{\mathbf{B}})$. The function d is a distance metric, which for all of our experiments is the standard Euclidian norm. The distribution of $cp([x, y, z], \mathbf{B})$ for voxels in $\mathcal{B}$ gives us a better sense of whether the voxels mistakenly labeled as hippocampus by our automatic algorithm are spatially near to, or far away from, the true location of the hippocampus. Likewise, the distribution of $cp([x, y, z], \hat{\mathbf{B}})$ for voxels in $\mathcal{C}$ gives us information about whether the voxels our automatic segmentation algorithm mistakenly labels as ”non-hippocampus” are close to, or distant from, the automatically estimated hippocampal surface.

For each subject image **I**, atlas-based segmentation provides an estimate of the hippocampi, and evaluating the quality of the estimate yields sets $\mathcal{B}$ and $\mathcal{C}$ of error voxels for **I**. Let $\mathcal{CP}_\mathbf{I}$ be the set of closest point distances for all error voxels in **I**, *i.e.* $\mathcal{CP}_\mathbf{I} = \{cp([x, y, z], \mathbf{B})|[x, y, z] \in \mathcal{B}\} \cup \{cp([x, y, z], \hat{\mathbf{B}})|[x, y, z] \in \mathcal{C}\}$. We compute one set $\mathcal{CP}_\mathbf{I}$ for each trial and wish to summarize CPDs over a set of trials– for example, the set of all trials on MCI subjects– into an interpretable statistic that summarizes the distribution of error voxels over all trials. To do so, we first compute a statistic– for example, the mean, median, or maximum– over each $\mathcal{CP}_\mathbf{I}$, then compute the mean of those statistics over all $\mathcal{CP}_\mathbf{I}$. For example, given a population of images, $\{\mathbf{I}1, \mathbf{I}2, \cdots\}$, we compute $mean(\{mean(\mathcal{CP}_{\mathbf{I}1}), mean(\mathcal{CP}_{\mathbf{I}2}), \cdots\})$. An alternative approach would be to pool all the CPDs over all images into a single set and compute statistics over that set, for example $mean(\mathcal{CP}_{\mathbf{I}1} \cup \mathcal{CP}_{\mathbf{I}2} \cdots)$. However, we feel that focusing on per-trial statistics provides a more intuitive sense of how the segmentation methods may perform for a particular atlas image, subject image, and registration method. In our results, we refer to the median CPD and maximum CPD for a particular $\hat{\mathbf{B}}$ and **B** as the *median error magnitude* and *maximum error magnitude* respectively.

4.5.3 Sectional analysis

Beyond computing overlap ratio and CPD measures over the entire hippocampus, we divide the hippocampus into sections and compute performance measures over voxels in each section. Doing so allows us to characterize the performance of our algorithms in terms of hippocampal sub-regions, which we feel is important for at least two reasons. First, certain portions of the hippocampus (for example, the head) may be more or

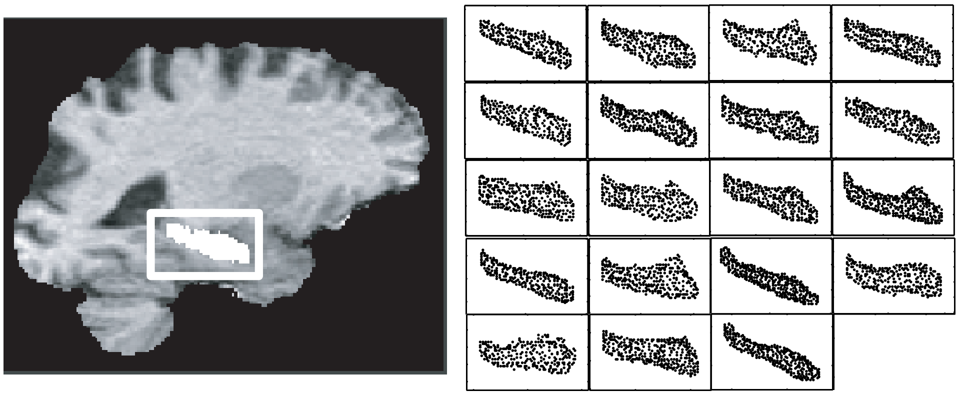

Figure 4: Points on the left hippocampus in all 19 MCI subjects are shown projected onto the XZ plane of the image. Note that all the hippocampi share the same rough initial orientation in this plane.

less important to segment accurately for some applications. Second, if atlas-based segmentation is used as an initial step in a larger segmentation pipeline, later steps in the pipeline (based on parametric shape models, for example [30]) could be optimized so that they focus computation on properly segmenting the hippocampal regions that were segmented poorly by the atlas-based step.

Consider a bounding box $(x_{min}, x_{max}, y_{min}, y_{max}, z_{min}, z_{max})$ around all the structure voxels in $\hat{\mathbf{B}}$ and $\mathbf{B}$ (*i.e.*, the x coordinates of all voxels in $\mathcal{A} \cup \mathcal{B} \cup \mathcal{C}$ are between x_{min} and x_{max}, etc.). For each of the three cardinal directions, we partition the estimated and ground-truth hippocampi into k sections along that direction and compute overlap ratios and CPDs in each of the sections. That is, for all i from 1 to k we compute $or(\mathbf{B}_i^x, \hat{\mathbf{B}}_i^x)$ and $\mathcal{CP}_{\mathbf{I}_i^x}$, where $\mathbf{B}_i^x(x, y, z) = \mathbf{B}(x, y, z)$ for $x_{min} + \frac{i-1}{k} * (x_{max} - x_{min}) < x < x_{min} + \frac{i}{k} * (x_{max} - x_{min})$ and $\mathbf{B}_i^x(x, y, z) = 0$ for all other voxels. Similarly, we compute $or(\mathbf{B}_i^y, \hat{\mathbf{B}}_i^y)$, $\mathcal{CP}_{\mathbf{I}_i^y}$, $or(\mathbf{B}_i^z, \hat{\mathbf{B}}_i^z)$ and $\mathcal{CP}_{\mathbf{I}_i^z}$ for all i from 1 to k. See Figure 3f for an illustration. In our experiments we set k to 10.

Figure 4 suggests that since the hippocampi all have similar orientations in the image, the sections can be interpreted as corresponding to rough anatomical regions on the hippocampus. For example, if we cut the shown hippocampi into sections using vertical lines as in Figure 3f, the sections to the left correspond roughly to posterior hippocampal regions, and sections to the right correspond to anterior regions. Likewise, cutting the hippocampi with horizontal lines divides the structures into sections that run from their inferior to superior extents respectively. This rough correspondence between axis-aligned sections and hippocampal regions allows us to meaningfully average the performance measures for the same section over many trials.

4.6 Statistical analysis: mixed-effects models

We analyze the effects of factors such as registration method, side of the brain, and disease state on segmentation performance measures through mixed-effects statistical models [39] that properly account for fixed effects, random effects, and grouping in our data. The fixed effects, including disease state, side of the brain, and registration method, are modeled as additive offsets from a baseline value of the performance measure. Random effects, such as the random sampling of subjects from an overall patient population, are modeled as variance components. Each level of each fixed effect is assigned a coefficient representing the offset it produces from the baseline value; for example, the fixed effect of disease state would be assigned three coefficients, corresponding to the additive contribution that being a control, MCI, or AD subject has on the dependent variable. We test for the overall significance of each fixed effect using Wald tests. Furthermore, we analyze differences between factor levels– for example, between control, MCI, and AD subjects– by using Wald tests to check for significant differences between their coefficients. In our analysis, between-group differences refer to differences in model coefficients between two factor levels. Mixed-effects models are important for our results for three main reasons. First, they properly account for the fact that we randomly sampled the subject images from overall populations of AD, MCI, and control subjects. Second, they model the random selection of cohort atlas images from a larger population. Third, the mixed-effects models account for repeated measures in our data; that is, the fact that we measure segmentation performance on the same subject images repeatedly for different factor levels. All statistics were performed using `R` version 1.9.1. Mixed-effects models were fit using maximum likelihood estimation in the `nlme` package.

5 Results

The following sections summarize the results of applying atlas-based segmentation techniques to the 54 images of AD, MCI, and control subjects. First, we explore the effects of registration method, disease state, and side of the brain on cohort-atlas-based segmentation (see Section 5.1). The effects of standard atlas, atlas mask, registration method, side of the brain, and disease state on standard-atlas-based segmentation performance are investigated in Section 5.2. Differences in performance measures between cohort-atlas-based and standard-atlas-based segmentation for a particular registration method are discussed in Section 5.3. Results comparing automated segmentation performance to manual-manual segmentation agreement are presented in Section 5.4. Additionally, we explore how the quality of segmentation varies across hippocampal sub-regions in Section 4.5.3. Results are discussed in more detail in Section 6.

5.1 Cohort atlases

For cohort-atlas-based segmentation, we fit mixed-effects models in which disease state, side of the brain, and registration method were fixed effects; the subject and cohort atlas identity were random effects; and the performance measures were the dependent variables. The overall effects of side, disease, and method on overlap ratio were statistically significant ($p < .0001$, $p = .0192$,$p < .0001$). The overall effects of side and method on maximum CPD were statistically significant ($p < .0001$, $p < .0001$), and the effects of side and method on median CPD were statistically sig-

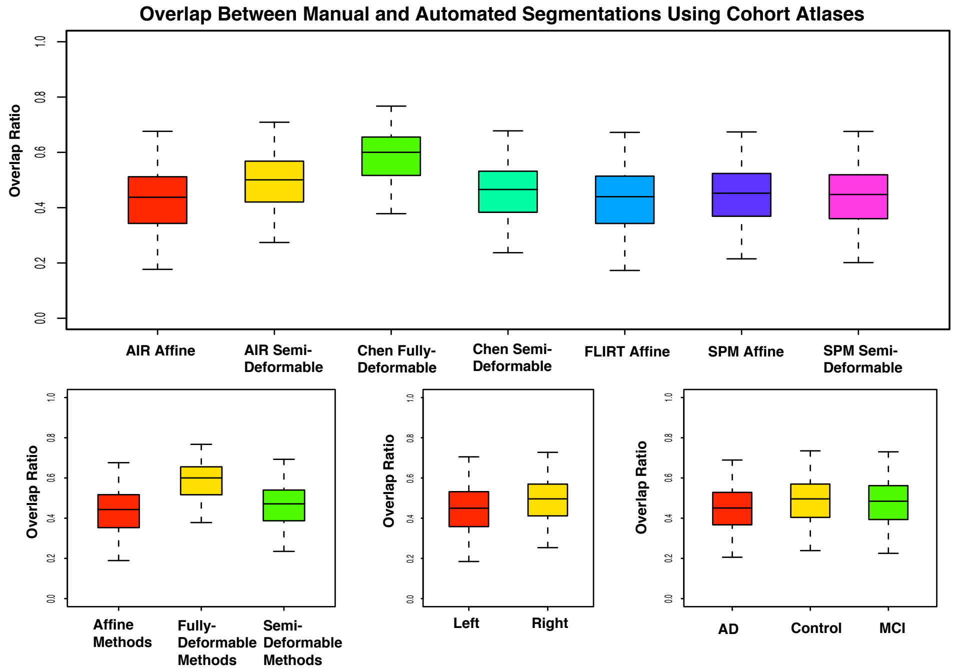

Figure 5: Overlap ratio as a function of disease state, registration method, and side of the brain for the 54 images using cohort atlases.

nificant ($p < .0001$, $p < .0001$). The effect of disease state on median CPD and maximum CPD were not statistically significant ($p = .959$ and $p = .412$ respectively).

Differences in model coefficients between individual registration methods, disease states, and sides of the brain were statistically analyzed. Overlap ratio was significantly lower in AD compared to MCI ($p = 0.0239$) and control ($p = .011$) groups. No significant difference in overlap ratio was seen between MCI and control groups ($p = .647$). No significant difference existed between the FLIRT affine and AIR affine methods ($p = .286$). For all other pairs of methods, significant (but in many cases slight) differences in overlap ratio existed ($p < .001$). The methods, ranked in decreasing order of overlap ratio, were as follows: Chen fully-deformable, AIR semi-deformable, Chen semi-deformable, SPM affine, SPM semi-deformable, FLIRT affine, AIR affine. In terms of median error magnitude, no significant difference existed between the Chen semi-deformable method and SPM affine methods ($p = .734$), or between the FLIRT affine and SPM semi-deformable methods ($p = .0545$). For all other pairs of methods, differences in median error magnitude were statistically significant. No significant differences in median error magnitude existed between AD and MCI ($p = .888$), AD and controls ($p = .872$), or MCI and controls ($p = .774$). The methods, ranked in increasing order of median error magnitude, were: Chen fully-deformable, AIR semi-deformable, Chen semi-deformable, SPM affine, SPM semi-deformable, AIR affine. No significant difference in maximum error magnitude ex-

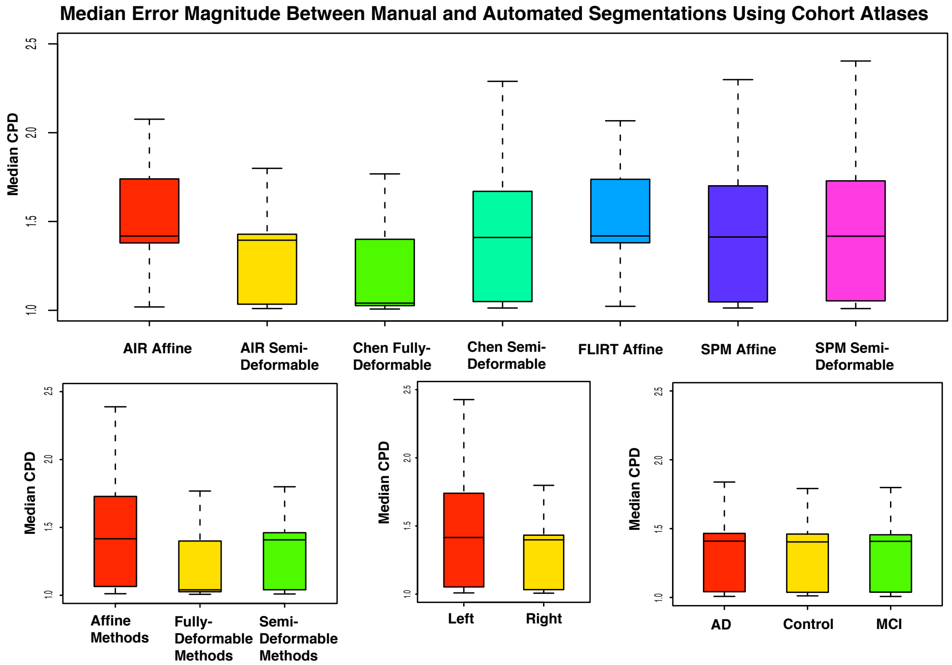

Figure 6: Median CPD as a function of disease state, registration method, and side of the brain for the 54 images using cohort atlases.

isted between AD and MCI ($p = .335$), AD and control ($p = .699$), or MCI and control ($p = .208$) groups. Furthermore, no significant difference existed between AIR affine and SPM semi-deformable ($p = .295$), FLIRT affine and SPM semi-deformable ($p = .087$), or AIR semi-deformable and Chen fully-deformable methods ($p = .133$). Differences between all other pairs of methods were significant in the model. The methods, ranked in increasing order of maximum error magnitude, were: Chen fully-deformable, AIR semi-deformable, SPM affine, FLIRT affine, SPM semi-deformable, AIR affine, Chen semi-deformable. Box plots showing how overlap ratio, median error magnitude, and maximum error magnitude vary with disease state, side of the brain, registration method, and registration method category, are shown in Figures 5, 6, and 7.

Comparing fully-deformable, semi-deformable and affine methods We grouped the registration methods into fully-deformable, semi-deformable, and affine categories (See Section 3.3.5) and fit a mixed-effects model in which the fixed effects were the method category, disease state, and side of the brain; subject and atlas identity were random effects. Fully-deformable methods had significantly higher overlap ratio and lower median and maximum error magnitudes than semi-deformable and affine methods ($p < .001$ in each case). In turn, semi-deformable methods had significantly higher overlap raio and lower median and maximum error magnitudes than affine methods

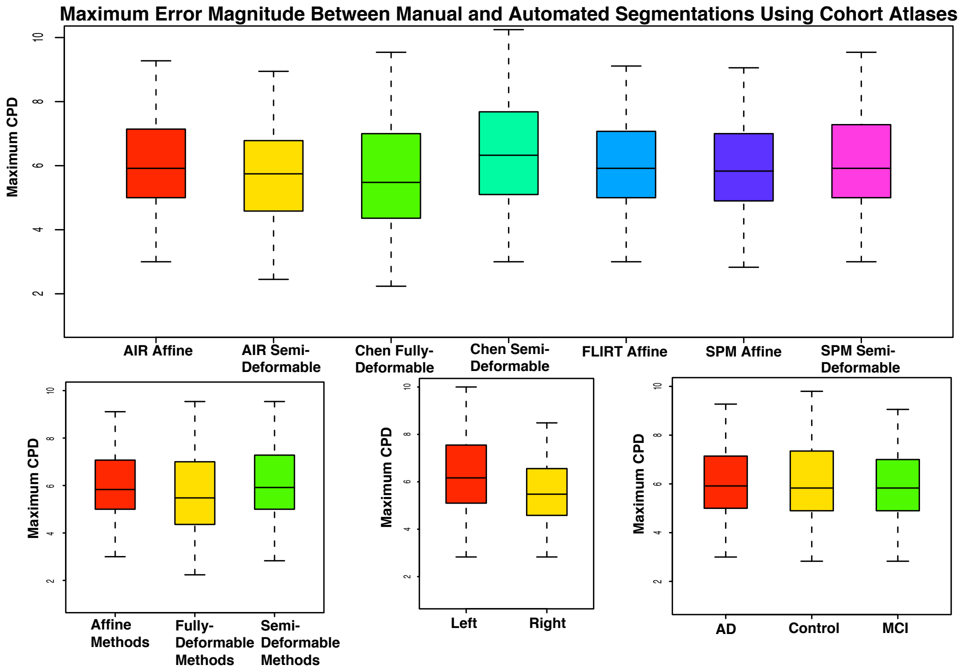

Figure 7: Maximum closest-point distance between automated cohort-atlas-based segmentations and manual segmentations for different registration methods, types of registration methods, sides of the brain, and disease states. See Section ??? for a description of significant differences between groups.

($p < .001$ in each case).

5.2 Standard atlases and atlas masks

For standard-atlas-based segmentation, we fit mixed-effects models in which the fixed effects were the atlas (Harvard vs. MNI), the source of the manual segmentation (R1 vs. Harvard/MNI), side of the brain, disease state, and registration method; subject identity was a random effect; and the performance measures were the dependent variables. Figures 8, 9, and 10 plot the overlap ratio, median error magnitude, and maximum error magnitude as a function of atlas image and atlas mask, registration method, side of the brain, and disease state. Results based on R1-traced atlas masks are referred to as ”Harvard By R1” and ”MNI By R1”; results based on atlas masks provided by the atlas institution are referred to as ”Harvard By Harvard” and ”MNI By MNI” respectively. Overlap ratio was significantly higher for R1-traced atlas hippocampi than hippocampi traced by the atlas institution ($p < .001$). No significant difference in overlap ratio was seen between the MNI and Harvard atlases ($p = .900$). Overlap ratio was significantly higher for right sides of the brain compared to left ($p < .001$). Overlap ratio was significantly lower for AD subjects than for MCI subjects ($p = .004$) and controls ($p = .020$), but no significant difference was seen between the MCI and control groups ($p = .665$). The registration methods, ranked in decreasing order of overlap ra-

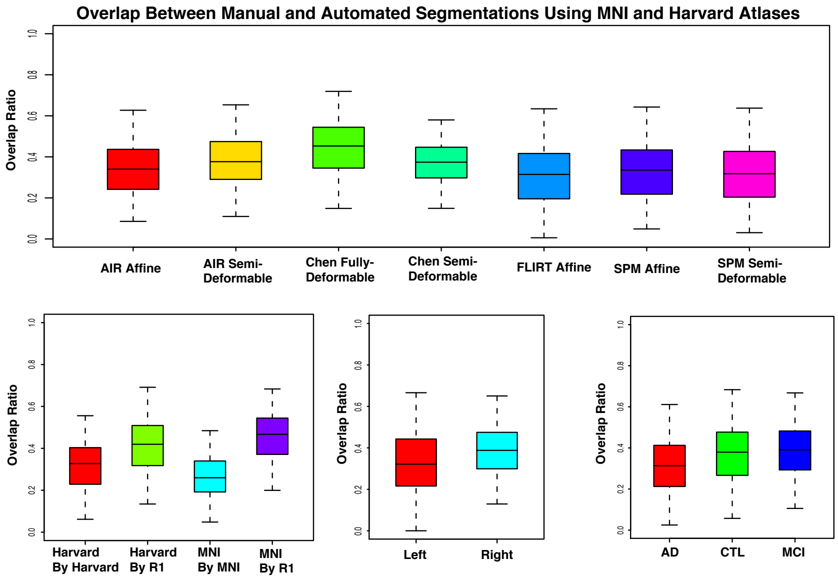

Figure 8: Overlap ratio as a function of disease state, registration method, and side of the brain for the 54 images using standard atlases.

tio, were: Chen fully-deformable, AIR semi-deformable, Chen semi-deformable, SPM affine, AIR affine, SPM semi-deformable, FLIRT affine. The difference in overlap ratio between the SPM semi-deformable and FLIRT affine methods was not statistically significant ($p = .163$), nor was the difference in overlap ratio between the Chen semi-deformable method and AIR semi-deformable method ($p = .072$). Differences in overlap ratio between all other pairs of methods were significant ($p < .05$).

Median error magnitude was significantly lower for R1-traced atlas hippocampi compared to hippocampi traced by the atlas institution ($p < .001$). No significant difference in median error magnitude was seen between the MNI and Harvard atlases ($p = .900$). Median error magnitude was significantly lower for right hippocampi compared to left ($p < .001$). No significant differences in median error magnitude were seen between AD and MCI($p = .258$), AD and control($p = .212$), or MCI and control ($p = .851$) groups. Furthermore, no significant differences in error magnitude were seen between any pairs of registration methods.

Maximum error magnitude is significantly lower in R1-traced atlas hippocampi than hippocampi traced by the atlas institution ($p < .001$). Significant differences in maximum error magnitude were seen between the MNI and Harvard atlases ($p < .001$). Differences in maximum CPD between left and right sides of the brain were significant ($p = .034$). Significant differences existed between AD and MCI ($p = .002$) and AD and control ($p = .010$) groups, but not between MCI and controls ($p = .679$).

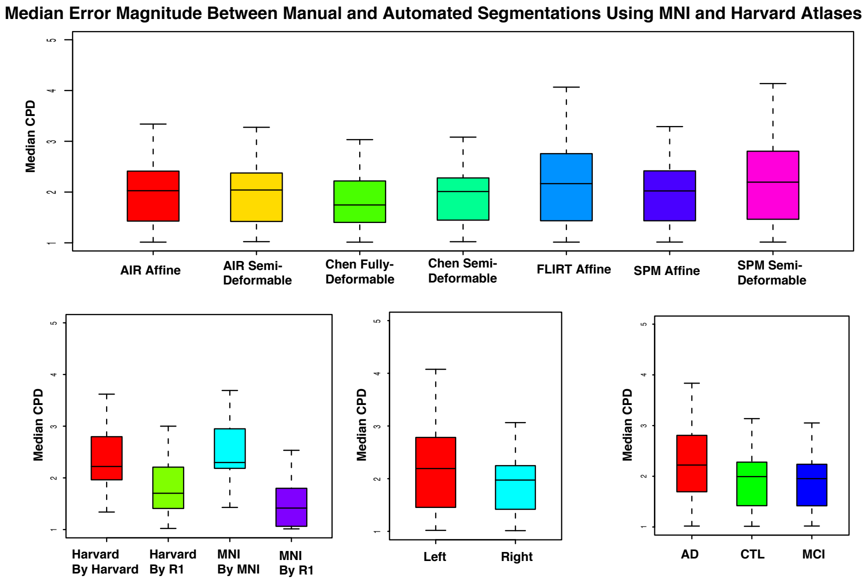

Figure 9: Median CPD as a function of disease state, registration method, and side of the brain for the 54 images using standard atlases.

Differences in maximum CPD were significant between the Chen semi-deformable method and all methods except the AIR affine method ($p < .028$). No significant differences in maximum CPD were seen between any other pair of methods.

5.3 Cohort atlases vs. standard atlases

We directly compared cohort-atlas-based segmentation to standard-atlas-based segmentation using the Chen fully-deformable registration method, which had shown the highest segmentation performance in experiments described in the previous sections. We fit mixed-effects models in which the atlas (MNI, Harvard, or cohort atlas), tracer (R1 or the atlas institution), side of the brain, and disease state were fixed effects, the subject identity was a random effect, and the dependent variables were the overlap ratio, median error magnitude, and maximum error magnitude. The mean overlap ratio was significantly higher for cohort-atlas-based segmentation than standard-atlas-based-segmentation using manual tracings by R1 along with the MNI ($p < .001$) or Harvard ($p < .003$) atlas images. Median error magnitude was significantly lower using cohort atlases than either standard atlas with manual tracings by R1 ($p < .001$ for MNI, $p < .001$ for Harvard). Maximum error magnitude was also significantly lower for cohort atlases than either standard atlas with manual tracings by R1 ($p < .001$ for MNI, $p < .001$ for Harvard). Performance measures for standard atlases using manual tracings from the atlas institution were significantly worse in each case. Figure

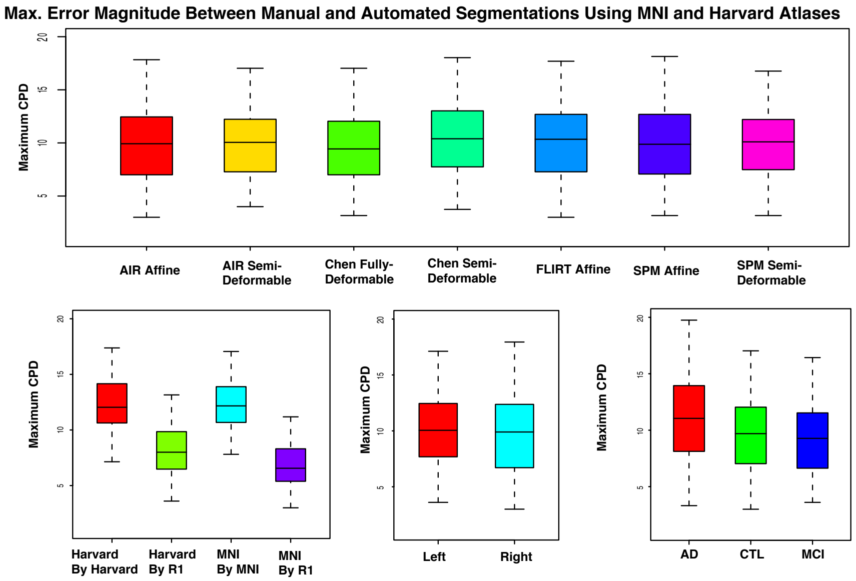

Figure 10: Maximum CPD as a function of disease state, registration method, and side of the brain for the 54 images using standard atlases.

11 plots performance measures between standard-atlas-based and cohort-atlas-based segmentation techniques.

5.4 Comparing manual-automated agreement to manual-manual agreement

Above, our statistical models measured the performance of automated segmentation algorithms in terms of *manual-automated agreement*, that is, agreement between automatic hippocampus segmentations and manual segmentations performed by an expert rater. Here, we compare manual-automated agreement to *manual-manual agreement*, or the agreement between manual segmentations performed by pairs of expert human raters. In so doing, we assess whether switching from manual to automated segmentation significantly increases the variability between the produced segmentation and one produced by an independent human rater. We selected 2 AD, 2 MCI, and 2 control images from our pool of subjects and had the hippocampi segmented manually by human raters R1, R2 and R3. Since rater R1 segmented hippocampi on the full set of 54 brains, we assess manual-automated agreement in terms of agreement between R1-rated manual segmentations and the Chen fully-deformable automated technique. Manual-manual agreement is measured in terms of pairwise agreement between manual segmentations by R1 and R2, R1 and R3, and R2 and R3. Manual-automated

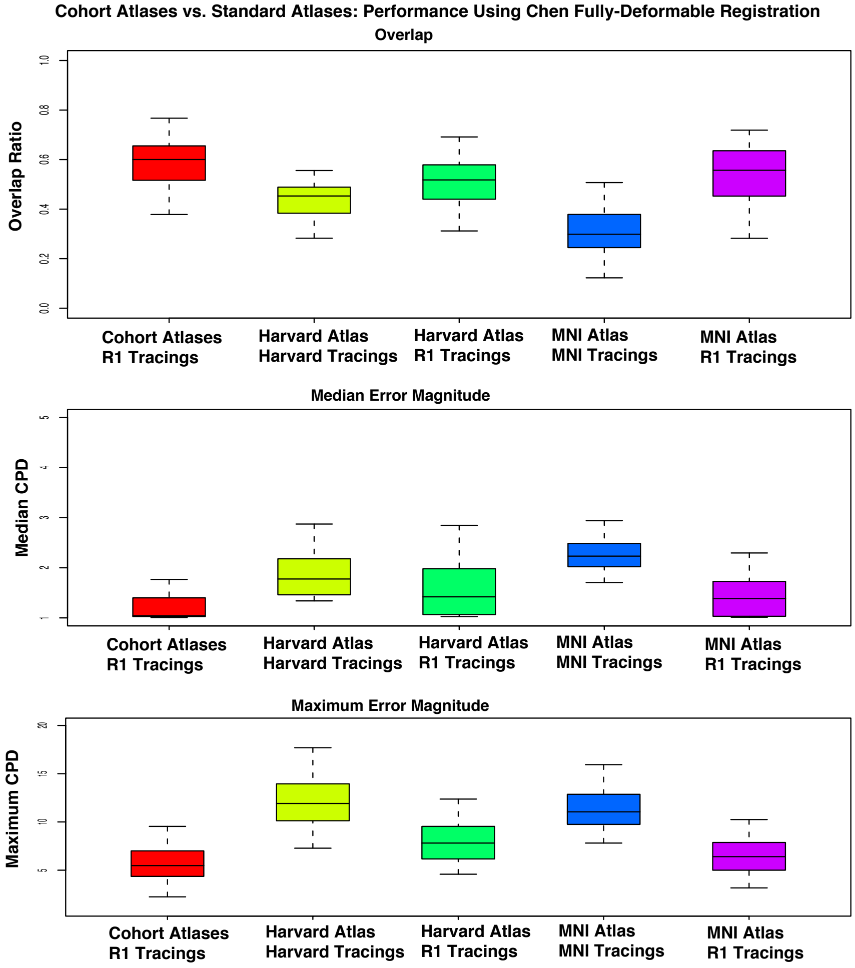

Figure 11: Overlap ratio, median CPD, and maximum CPD between cohort-atlas-based and standard-atlas-based segmentation.

agreement for each subject is measured in terms of the average agreement between its R1 segmentation and the automated segmentations from all cohort atlases in its disease category. We quantify manual-manual and manual-automated agreement on a per-hippocampus basis in terms of the performance measures described in Section 4.5– that is, for a pair of segmentations of the same hippocampus (performed by R1, R2, R3, or the automated technique, respectively), we quantify agreement between segmentations in terms of the overlap ratio, median error magnitude, and maximum error magnitude. We fit mixed-effect models with the agreement measures as dependent variables, the type of agreement (manual-manual or manual-automated) and side of the brain as fixed effects, and subject identity as a random effect. Note that this

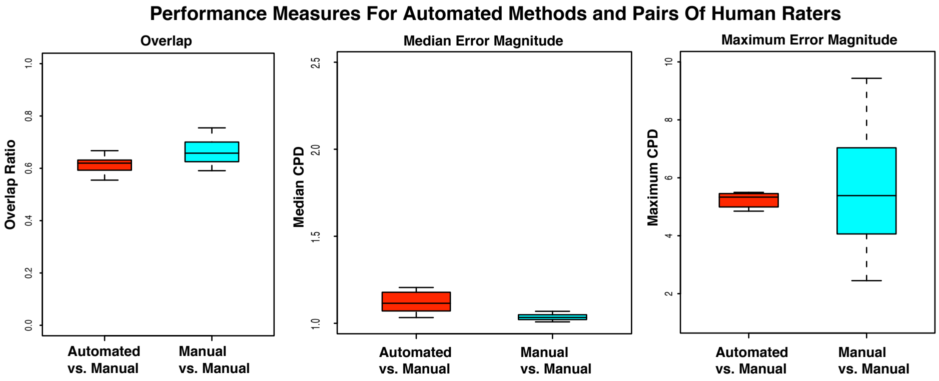

Figure 12: Overlap ratio, median CPD, and maximum CPD between manual and automated segmentations (automatic vs. manual) and between pairs of manual segmentations (manual vs. manual).

approach differs from the more common approach of measuring agreement between pairs of raters in terms of hippocampal volumes; the key difference is that our approach quantifies agreement in terms of how well the segmentations overlap in the brain. Other approaches, based on estimating automated segmentation performance and the true, underlying structure mask simultaneously, are also available [52]. Manual-manual agreement was not significantly higher than manual-automated agreement in terms of overlap ratio ($p = .0916$). Furthermore, differences in median and maximum error magnitude were not significant between manual-manual agreement and manual-automated agreement ($p = .775$ and $p = .455$ respectively). Box plots comparing the distribution of agreement measures for manual-manual and manual-automated agreement are shown in Figure 12.

5.5 Sectional results

Figures 13, 14, and 15 plot mean overlap ratios for hippocampal sections along the three cardinal directions of our data set. The three cardinal directions correspond roughly to the posterior-anterior, medial-lateral, and superior-anterior hippocampal axes, respectively (see Section 4.5.3 and Figure 4). For all methods, the hippocampal sections most responsible for segmentation error are located at the extremities of the hippocampus, especially at the superior, inferior, medial, and lateral ends. With the exception of the most extreme sections, mean overlap ratio is generally higher toward the lateral extent of the hippocampus and lower toward the medial extent (Figure 13). Furthermore, with the exception of the most extreme sections, mean overlap ratio is relatively constant with respect to anterior-posterior position (Figure 14). Finally, moving from the superior to inferior extent, mean overlap ratio increases steadily, reaches a peak at the central sections, and decreases toward the inferior end (Figure 15). These distributions of mean overlap ratio do not vary significantly with respect to disease state, side of the brain, or registration method. In particular, it does not appear that the registration techniques vary significantly with respect to how overlap is distributed

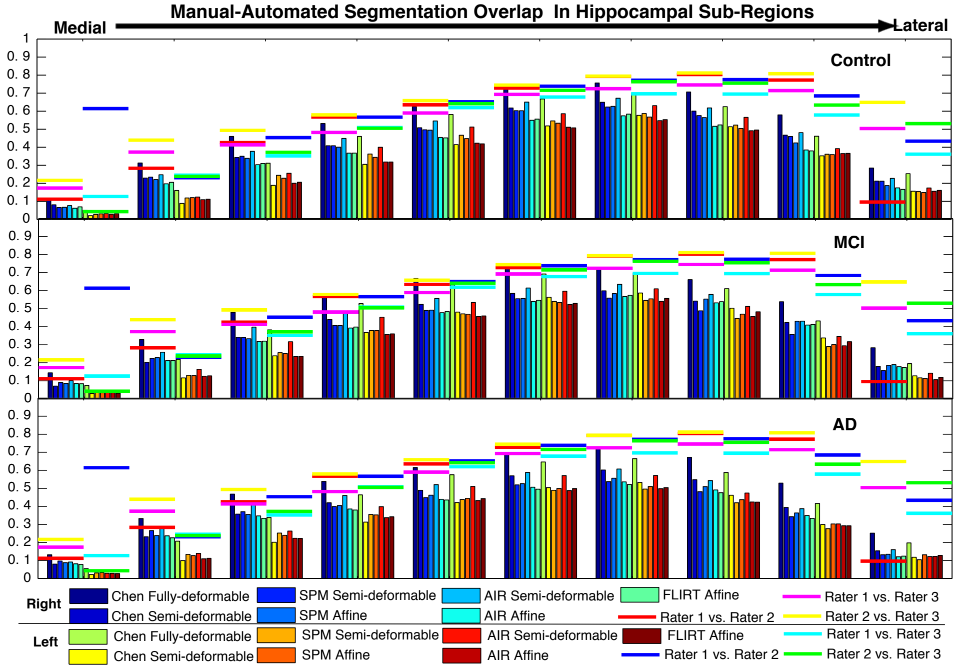

Figure 13: Overlap distance measures broken down along medial-lateral line for automatic registration methods and manual raters on control, MCI, and AD images. See text for details.

across hippocampal sub-regions. Furthermore, these patterns of manual-automated overlap across sub-regions are similar to patterns of manual-manual overlap on the 6 selected images, although the human raters are relatively more consistent at the lateral extent.

Figures 13, 14, and 15 plot the mean median error magnitude and mean maximum error magnitude across hippocampal sections. The mean median error magnitude is higher at the extremities, especially at the medial, posterior, and superior extents. However, the mean maximum error magnitude does not vary significantly with respect to the position of the hippocampal sub-region, although the mean maximum error magnitude is slightly higher at the posterior (Figure 14) and medial (Figure 15) extremities. The automated methods are largely competitive with manual raters in terms of mean median CPD at the central hippocampal sections; mean maximum error magnitude is competitive with human raters in medial, superior, and anterior sections. Interestingly, while the distributions of manual-automated mean overlap ratio are highly similar to the distributions of manual-manual mean overlap ratio, the corresponding patterns of mean median error magnitude and mean maximum error magnitude differ significantly. In particular, error voxels for pairs of manual raters are markedly closer to the hippocampal surface at the lateral extent, at the central sections along the posterior-anterior axis, and at the inferior extent. The fact that human raters are able to more consistently segment those sub-regions suggests that it would be possible to optimize automated methods to segment these sub-regions more effectively.

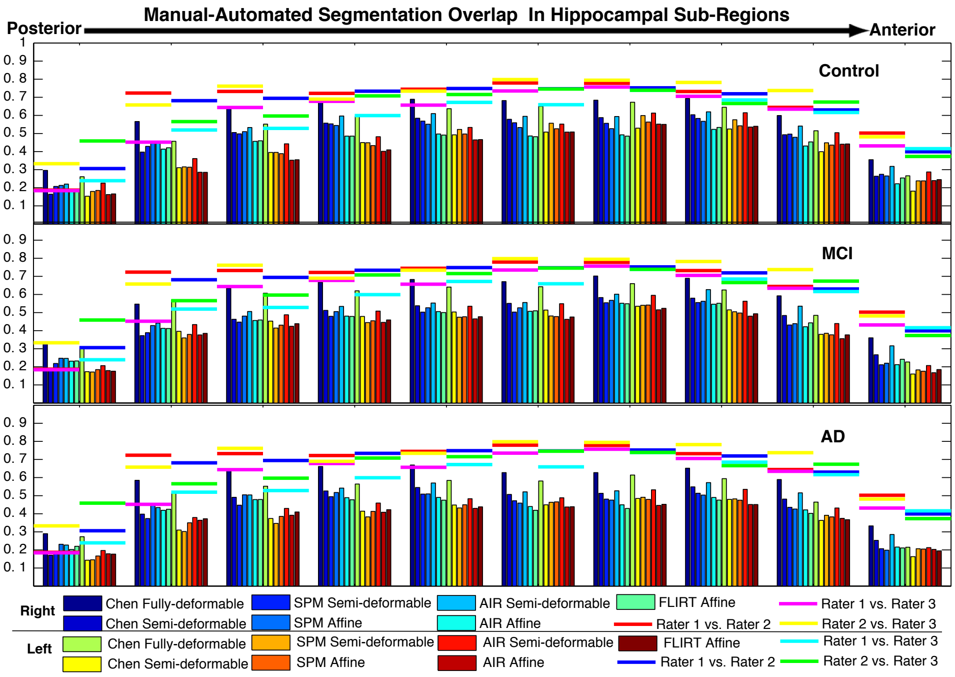

Figure 14: Overlap distance measures broken down along posterior-anterior line for automatic registration methods and manual raters on control, MCI, and AD images. See text for details.

6 Discussion

This section summarizes our results in terms of which factors led to higher or lower performance measures in the atlas-based segmentation experiments. A ”>” between two factor levels indicates that the overlap was higher, and/or the error magnitudes were lower, for the first factor level compared to the second.

Fully-deformable $>$ semi-deformable $\geq$ affine Our results confirm the intuition that methods making use of more highly-deformable geometric transformation models tended to be able to fit the complex shape of the hippocampus more accurately than less-deformable geometric models. We believe that the AIR semi-deformable technique performed better than competing semi-deformable methods because the ”deformability” of its geometric transformation– *i.e.*, the degree of its polynomial basis– was allowed to gradually increase over the course of optimization, while the geometric transformations for the Chen and SPM semi-deformable techniques were fixed in their spatial structure. Furthermore, as mentioned above, SPM is explicitly biased toward minimally-deforming transformations, which may steer its geometric transformation away from highly-accurate fit of the hippocampal surface.

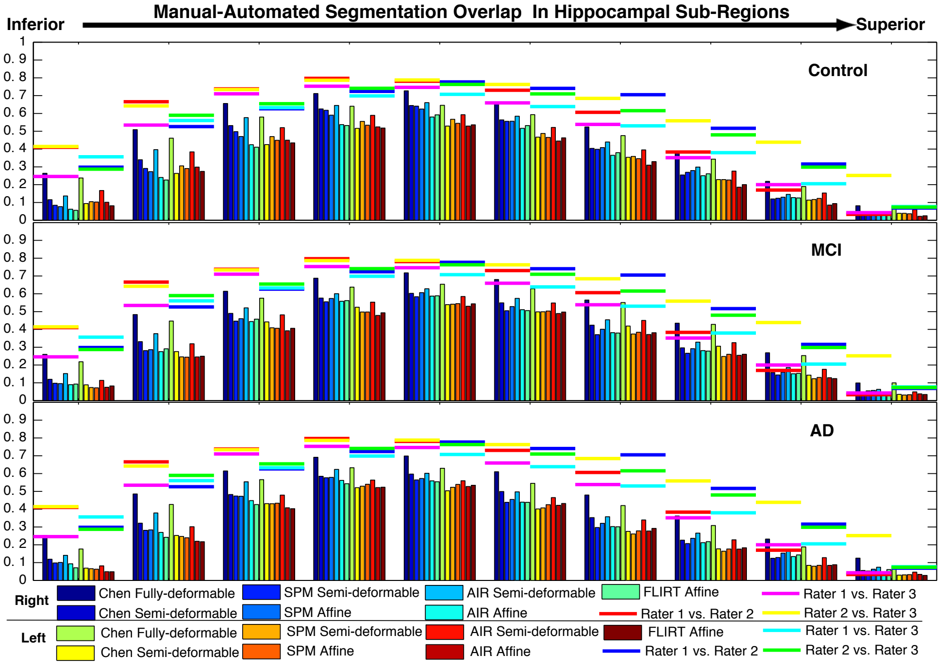

Figure 15: Overlap distance measures broken down along superior-anterior line for automatic registration methods and manual raters on control, MCI, and AD images. See text for details.

Human-human agreement ≈ automated-human agreement for fully-deformable registration Results suggest a general trend toward higher manual-manual agreement compared to manual-automated agreement (see Figures 13 and 12), but the differences are not statistically significant. Thus, while there may be room for improvement of the automated methods, Chen's fully-deformable method can be competitive with the human raters in terms of overlap, median error magnitude, and maximum error magnitude. Automated methods may be competitive for elderly hippocampus segmentation applications, especially those that can tolerate minor errors in the spatial localization of the hippocampus.

MCI ≈ controls > AD Overall performance measures are significantly lower among AD subjects than MCI or control subjects. One possible explanation for these results is that the degenerative proccesses of AD make image registration inherently more difficult and ambiguous by reducing tissue contrast and/or inducing a high degree of variability in the geometric characteristics of brain structures such as the hippocampus. Another possible explanation is that registering pairs of AD images is no more or less difficult than registering MCI or elderly control brains, but that standard software packages are not optimized for the task. Similarly, the fact that overlap ratios for MCI and control cases are similar could suggest that their image characteristics do not differ so significantly that they affect registration.

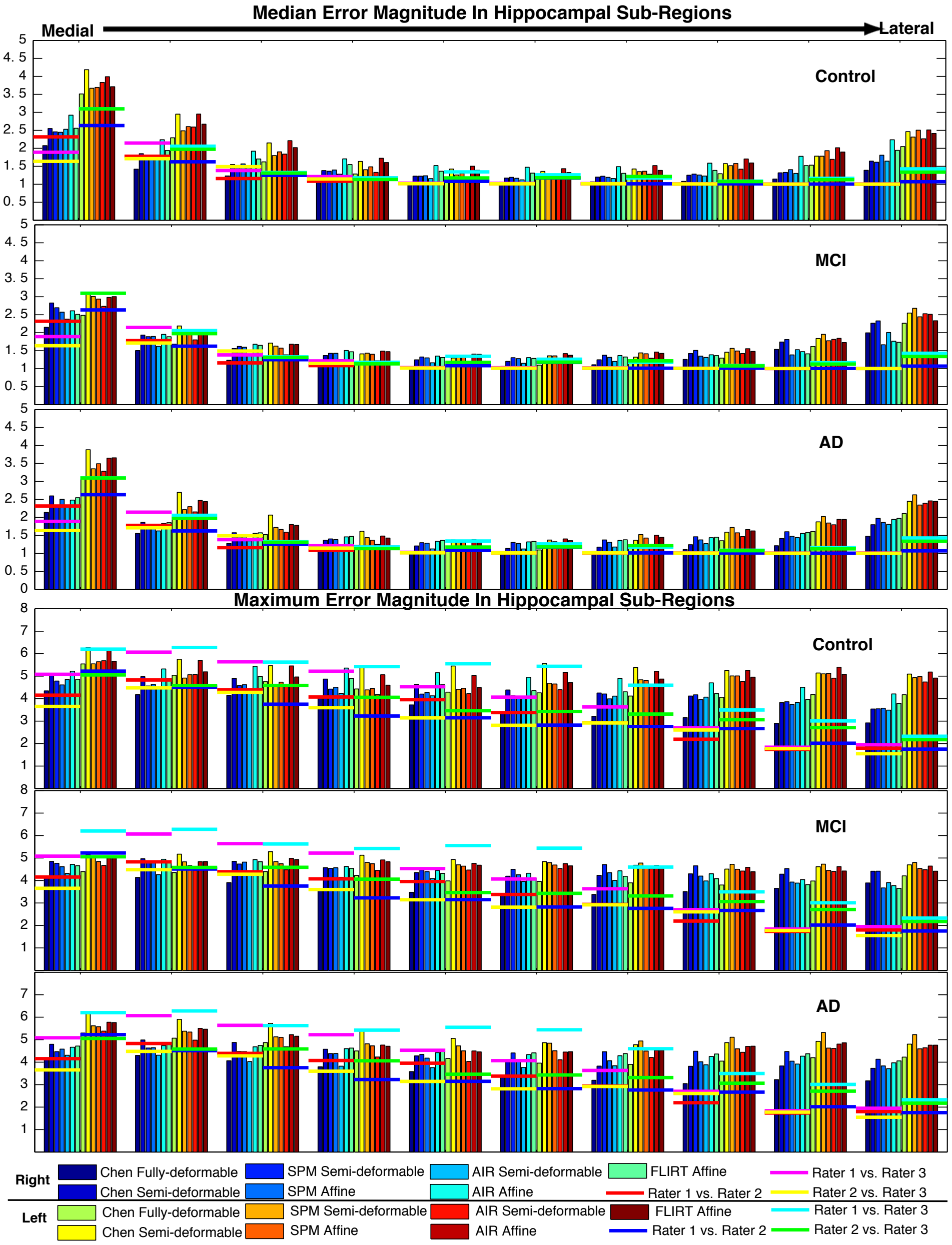

Figure 16: Closest-point distance measures broken down along medial-lateral line for automatic registration methods and manual raters on control, MCI, and AD images. See text for details.

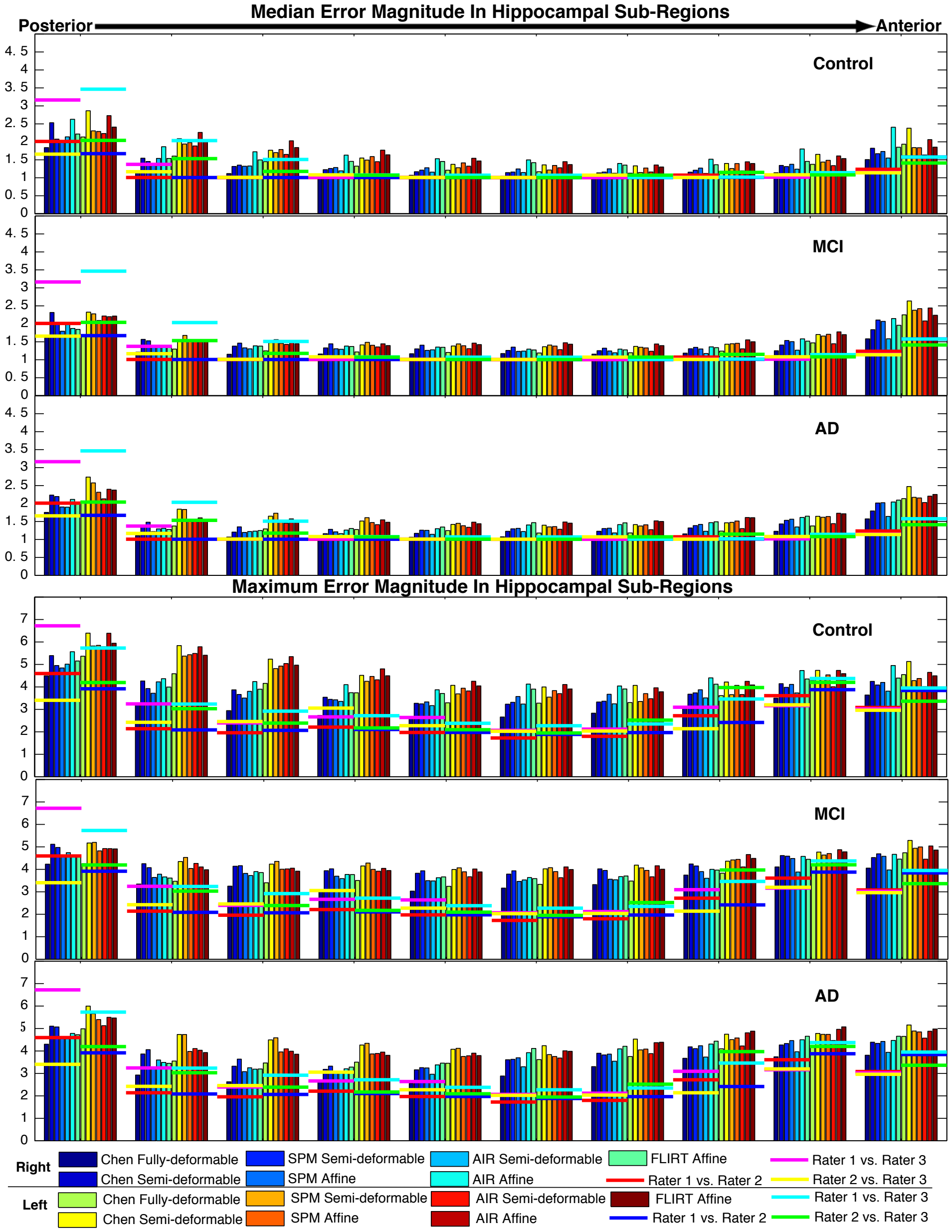

Figure 17: Closest-point distance measures broken down along posterior-anterior line for automatic registration methods and manual raters on control, MCI, and AD images. See text for details.

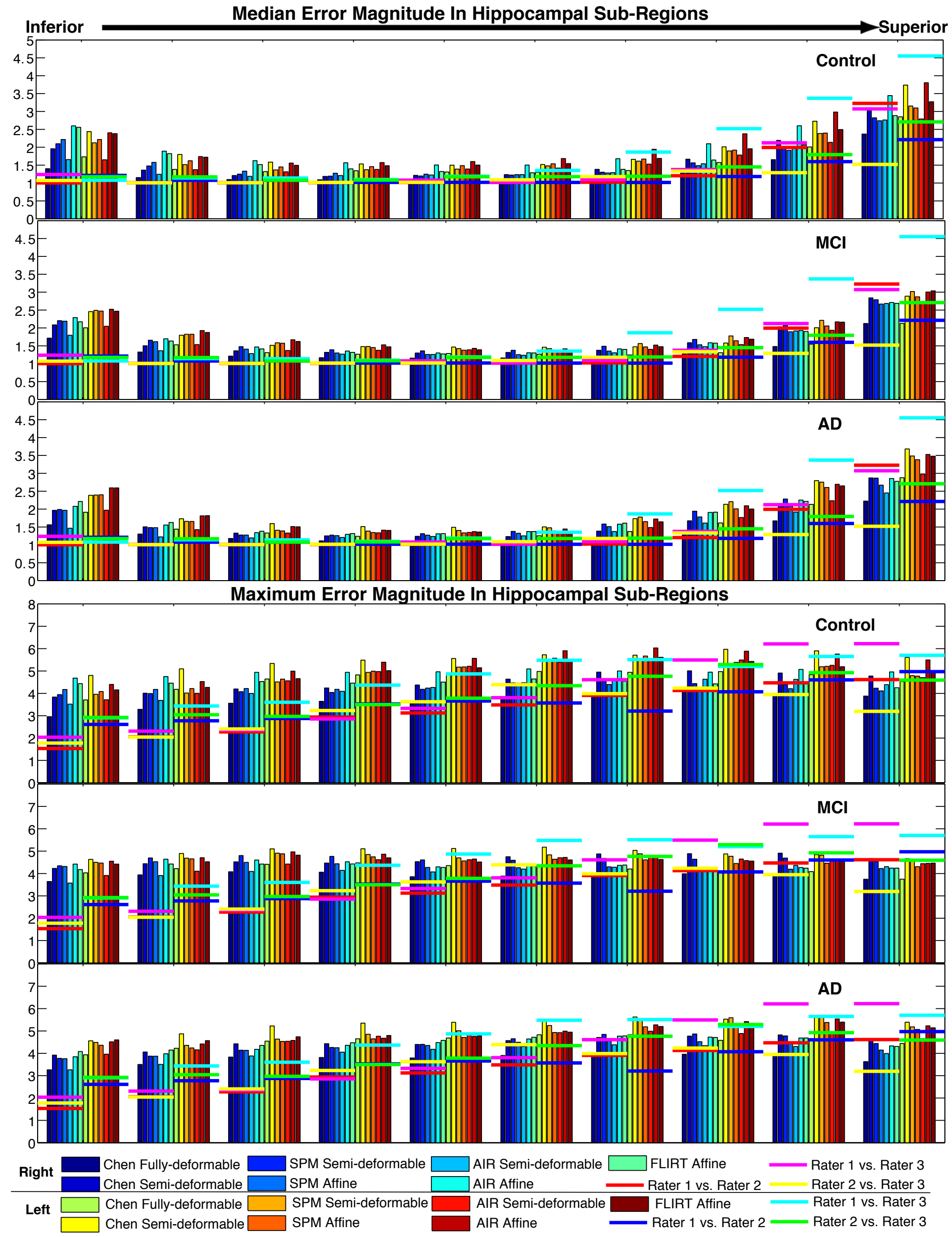

Figure 18: Closest-point distance measures broken down along superior-anterior line for automatic registration methods and manual raters on control, MCI, and AD images. See text for details.

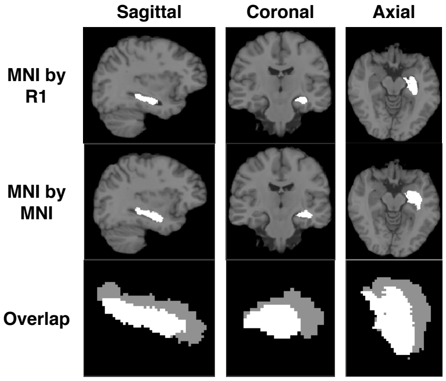

Figure 19: Comparison of manual hippocampus segmentations on the MNI atlas provided by MNI and Rater R1. In the bottom row, points in white are in overlap between the two tracings, and points in gray are in the MNI tracing only. Note that the MNI tracing is significantly larger than that by R1.

Right $>$ left A striking bilateral asymmetry in automated segmentation performance measures is seen in all experiments, across all three disease groups. These results echo the slight bilateral asymmetry in atlas-based hippocampus segmentation results shown by Duchesne *et al.* [17]. However, a mixed-effects model fit to solely manual-manual agreement data shows no significant bilateral asymmetry in manual-manual overlap ratio ($p = 0.12$), median error magnitude ($p = 0.0681$), or maximum error magnitude ($p = 0.6811$). Our initial calculations of hippocampal volumes do not show a significant volume asymmetry, echoing the findings of Bigler *et al.* [2], but it is possible that age- and AD-related atrophy has caused other morphological changes, such as hippocampal shape deformation or decreased tissue contrast, on the left side of the brain. These asymmetric changes could confound automated techniques in a way that expert human raters were able to compensate for. However, further investigation is needed to explain this bilateral effect.

Cohort-atlas-based $\geq$ standard-atlas-based Results from our mixed-effects models suggest that randomly selecting cohort atlas images from a population leads to higher automated segmentation performance than standard-atlas-based segmentation, independent of differences in manual segmentation protocols between institutions. This confirms our intuition that differences in brain morphology and image acquisition characteristics between atlas and subject images can negatively impact performance of atlas-based segmentation. In particular, differences in brain structure between the young, healthy individuals scanned for standard atlas images and the elderly subjects in our study could pose additional challenges to accurate image registration and segmentation. Future work should investigate the ways in which discrepancies in morphology, image acquisition parameters, and scanning equipment impact atlas-based segmentation results.

Posterior $\approx$ anterior, lateral $>$ medial, center $>$ periphery The sectional results presented in Section 4.5.3 suggest that segmentation errors are evenly distributed be-

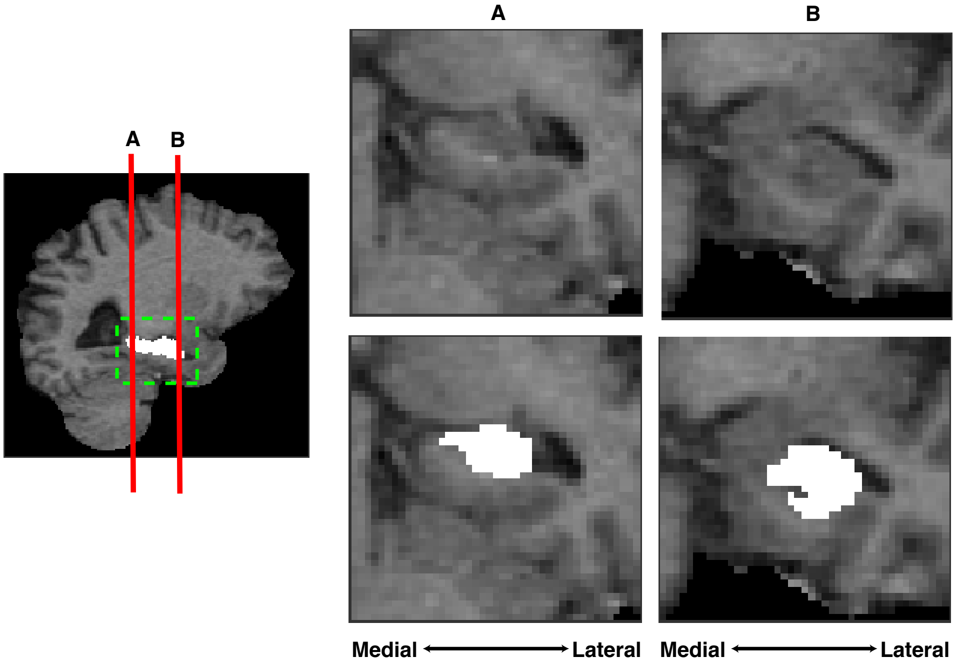

Figure 20: Two coronal slices from an example MCI image show areas where the hippocampus is bounded by gray matter and CSF on the lateral side, and entirely by gray matter on the medial side. Coronal slices A and B are shown magnified to show the region denoted by the green box. Voxels labeled as belonging to the hippocampus by rater R1 are shown in white in the bottom row. Note that hippocampus-CSF boundaries are relatively sharp and distinct, while interfaces with other gray matter structures are relatively subtle.

tween posterior and anterior regions of the hippocampus, are more concentrated in the medial regions than the lateral regions, and are generally more highly concentrated toward the periphery than the center. One possible reason for the medial skew in errors is that CSF forms part of the lateral boundary of the structure over its entire anterior-posterior extent, while in some regions, the medial boundary consists entirely of subtle, ambiguous interfaces with other gray-matter compartments (See Figure 20). We suggest that the sharp contrast between gray matter and CSF forms a strong visual cue that the automated methods take advantage of to more accurately localize the lateral boundary. Interestingly, our finding that agreement between pairs of human raters does not vary significantly along the anterior-posterior direction except at the extreme periphery contrasts with the inter-rater consistency maps shown by Thompson *et al.* [49], which suggest that manual tracings are relatively more consistent in the anterior sections. A possible explanation for this discrepancy is that the consistency measure of Thompson *et al.* is based on agreement between raters in radial distances from the medial axis of the hippocampus to its surface, and therefore could be more sensitive in posterior sub-regions where the radial distances are relatively small.

Manual tracing protocols add significant variability Geuze *et al.* recently described a dizzying array of existing protocols for manually segmenting the hippocampus in MR [22]. Our results (see Figure 8 indicate that discrepancies between these manual protocols can add a highly significant source of variation to what portion of the brain can be expected to be labeled as hippocampus, both in manual segmentation and atlas-based automated methods. Figure 19 gives an example of the significant discrepancies between manual segmentations of the MNI atlas image produced by R1 and by MNI. We emphasize that we are not suggesting that the manual segmentation protocol used by R1 is superior or inferior to those employed for the Harvard or MNI atlases; rather, we have showed that variations in the resulting hippocampi can be significant. Therefore, we suggest that researchers using standard atlas images for atlas-based segmentation should examine the atlas masks and tracing protocols closely to be sure the delineation conventions employed match those of their own laboratory. If they do not, our results have shown that tracing the structure on the standard atlas or on a randomly-selected subject image leads to automated segmentations whose agreement with expert manual segmentations is competitive with manual-manual agreement.

7 Conclusion

Atlas-based segmentation is a simple, automated method for hippocampus segmentation that can use standard image registration techniques to produce reasonable structure delineations in images of elderly controls and subjects with MCI and AD. While additional work is needed to make these automated techniques truly competitive with expert human raters, their performance may be acceptable for image proccessing applications that can tolerate a small amount of hippocampus localization error. While standard digital atlases from MNI, Harvard, and other institutions allow investigators to apply atlas-based segmentation to their subject images with no need for manual labeling, care must be taken to insure that hippocampus tracing protocols from the atlas institution coincide with those of the investigator.

References

[1] J Ashburner and KJ. Friston. Voxel-based morphometry–the methods. *NeuroImage*, 11(6):805–821, June 2000.

[2] Erin D. Bigler, David F. Tate, Michael J. Miller, Sara A. Rice, Cory D. Hessel, Heath D. Earl, Joann T. Tschanz, Brenda Plassman, and Kathleen A. Welsh-Bohmer. Dementia, asymmetry of temporal lobe structures, and apolipoprotein e genotype: Relationships to cerebral atrophy and neuropsychological impairment. *Journal of the International Neuropsychological Society*, 8:925–933, 2002.

[3] M Bobinski, J Wegiel, HM Wisniewski, M Tarnawski, M Bobinski, B Reisberg, MJ De Leon, and DC Miller. Neurofibrillary pathology–correlation with hippocampal formation atrophy in alzheimer disease. *Neurobiology of Aging*, 17(6):909–919, 1996.

[4] G Bueno, O Musse, F Heitz, and JP. Armspach. Three-dimensional segmentation of anatomical structures in mr images on large data bases. *Magnetic Resonance Imaging*, 19(1):73–88, January 2001.

[5] DT Chard, GJ Parker, CM Griffin, AJ Thompson, and DH. Miller. The reproducibility and sensitivity of brain tissue volume measurements derived from an spm-based segmentation methodology. *Journal of Magnetic Resonance Imaging*, 15(3):259–267, March 2002.

[6] Mei Chen. *3-D Deformable Registration Using a Statistical Atlas with Applications in Medicine*. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, October 1999.

[7] Gael Chetelat and Jean-Claude Baron. Early diagnosis of alzheimer's disease: contribution of structural neuroimaging. *Neuroimage*, 18:525–541, 2003.

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

[9] M.K. Chung, K.J. Worsley, T. Paus., C. Cherif, D.L. Collins, J.N. Giedd, J.L. Rapoport, and A.C. Evans. A unified statistical approach to deformation-based morphometry. *NeuroImage*, 14:595–606, 2001.

[10] D. Collins, C. Holmes, T. Peters, , and A. Evans. Automatic 3d model-based neuroanatomical segmentation. *Human Brain Mapping*, 3(3):190–208, 1995.

[11] D. L. Collins, T. M. Peters, W. Dai, , and A. C. (1992) Evans. Model based segmentation of individual brain structures from mri data. In *SPIE Vol. 1808, Visualization in Biomedical Computing*, pages 10–23, 1992.

[12] W.R. Crum, R.I. Scahill, and N.C. Fox. Automated hippocampal segmentation by regional fluid registration of serial mri: validation and application in alzheimer's disease. *NeuroImage*, 13(5):847–855, 2001.

[13] BM Dawant, SL Hartmann, JP Thirion, Maes F, D Vandermeulen, and P. Demaerel. Automatic 3-d segmentation of internal structures of the head in mr images using a combination of similarity and free-form transformations: Part i, methodology and validation on normal subjects. *IEEE Transactions on Medical Imaging*, 18(10):909–916, October 1999.

[14] R Perez de Alejo, J Ruiz-Cabello, M Cortijo, I Rodriguez, I Echave, J Regadera, J Arrazola, P Aviles, P Barreiro, D Gargallo, and M Grana. Computer-assisted enhanced volumetric segmentation magnetic resonance imaging data using a mixture of artificial neural networks. *Magnetic Resonance Imaging*, 21(8):901–912, October 2003.

[15] MJ de Leon, J Golomb, AE George, A Convit, CY Tarshish, T McRae, S De Santi, G Smith, SH Ferris, and M Noz et al. The radiologic prediction of alzheimer disease: the atrophic hippocampal formation. *American Journal of Neuroradiology*, 14(4):897–906, July-August 1993.

[16] B.C. Dickerson, D. H. Salat, J. F. Bates, M. Atiya, R. J. Killiany, D. N. Greve, A. M. Dale, C. E. Stern, D. Blacker, M. S. Albert, and R. A. Sperling. Medial temporal lobe function and structure in mild cognitive impairment. *Annals of Neurology*, 56(1):27–35, 2004.

[17] S. Duchesne, J.C. Pruessner, and D.L. Collins. An appearance-based method for the segmentation of medial temporal lobe structures. *NeuroImage*, 17(2):515–531, 2002.

[18] B Fischl, DH Salat, E Busa, M Albert, M Dieterich, C Haselgrove, A van der Kouwe, R Killiany, D Kennedy, S Klaveness, A Montillo, N Makris, B Rosen, and AM Dale. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33:341–355, 2002.

[19] PA Freeborough, Fox NC, and RI Kitney. Interactive algorithms for the segmentation and quantitation of 3-d mri brain scans. *Computer Methods and Programs in Biomedicine*, 53(1):15–25, May 1997.

[20] GB Frisoni. Structural imaging in the clinical diagnosis of alzheimer's disease: problems and tools. *Journal of Neurology, Neurosurgery, and Psychiatry*, 70(6):711–718, September 2001.

[21] K.J. Friston, J. Ashburner, C.D. Frith, J.-B. Poline, J.D. Heather, , and R.S.J. Frackowiak. Spatial registration and normalization of images. In *Annual Meeting of the Organization for Human Brain Mapping*, pages 165–189, 1995.

[22] E Geuze, E Vermetten, and JD Bremner. Mr-based in vivo hippocampal volumetrics: 1. review of methodologies currently employed. *Molecular Psychiatry*, pages 1–13, August 31 2004.

[23] J. W. Haller, A. Banerjee, G. E. Christensen, M. Gado, S. Joshi, M. I. Miller, Y. Sheline, M. W. Vannier, , and J. G. Csernansky. Three-dimensional hippocampal mr morphometry with high-dimensional transformation of a neuroanatomic atlas. *Radiology*, 202:504–510, 1997.

[24] Steven L. Hartmann, Mitchell H. Parks, Peter R. Martin, and Benoit M. Dawant. Automatic 3-d segmentation of internal structures of the head in mr images using a combination of similarity and free-form transformations: Part ii, validation on severely atrophied brains. *IEEE Transactions on Medical Imaging*, 18(10):917–926, October 1999.

[25] Robert E. Hogan, Kevin E. Mark, Lei Wang, Sarang Joshi, Michael I. Miller, and Richard D. Bucholz. Mesial temporal sclerosis and temporal lobe epilepsy: Mr imaging deformation-based segmentation of the hippocampus in five patients. *Radiology*, 216:291–297, 2000.

[26] D.V. Iosifescu, M.E. Shenton, S.K. Warfield, R. Kikinis, J. Dengler, F.A. Jolesz, and R.W. McCarley. An automated registration algorithm for measuring mri subcortical brain structures. *NeuroImage*, 6(1):13–25, 1997.

[27] CR Jack, MD Bentley, CK Twomey, and AR Zinsmeister. Mr imaging-based volume measurements of the hippocampal formation and anterior temporal lobe: validation studies. *Radiology*, 176:205–209, 1990.

[28] Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. Improved methods for the registration and motion correction of brain images. Technical report, Oxford Centre for Functional Magnetic Resonance Imaging of the Brain, 2002.

[29] CR Jack Jr, WH Theodore, M Cook, and G McCarthy. Mri based hippocampal volumetrics: data acquisition, normal ranges, and optimal protocol. *Magnetic Resonance Imaging*, 13:1057–1064, 1995.

[30] A Kelemen, G Szekely, and G Gerig. Elastic model-based segmentation of 3-d neuroradiological data sets. *IEEE Transactions on Medical Imaging*, 18(10):828–839, October 1999.

[31] Ron Kikinis, Chiara M. Portas, Robert M. Donnino, Ferenc A. Jolesz, Martha E. Shenton, Dan V. Iosifescu, Robert W. McCarley, Pairash Saiviroonporn, Hiroto H. Hokama, Andre Robatino, David Metcalf, and Cynthia G. Wible. 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.

[32] Jan Klemencic, Vojko Valencic, and Nuska Pecaric. Deformable contour based algorithm for segmentation of the hippocampus from mri. In *Proceedings of the International Conference on Computer Analysis of Images and Patterns*, pages 298–308, 2001.

[33] JH Kordower, Y Chu, GT Stebbins, ST DeKosky, EJ Cochran, D Bennett, and EJ Mufson. Loss and atrophy of layer ii entorhinal cortex neurons in elderly people with mild cognitive impairment. *Annals of Neurology*, 49(2):202–213, February 2001.

[34] KV Leemput, F Maes, D Vandermeulen, and P. Suetens. Automated model-based tissue classification of mr images of the brain. *IEEE Transactions on Medical Imaging*, 18:897–908, 1999.

[35] OL Lopez, JT Becker, W Klunk, J Saxton, RL Hamilton, DI Kaufer, R Sweet, C Cidis Meltzer, S Wisniewski, MI Kamboh, and ST DeKosky. Research evaluation and diagnosis of probable alzheimer's disease over the last two decades: I. *Neurology*, 55:1854–1862, 2000.

[36] OL Lopez, JT Becker, W Klunk, J Saxton, RL Hamilton, DI Kaufer, R Sweet, C Cidis Meltzer, S Wisniewski, MI Kamboh, and ST DeKosky. Research evaluation and diagnosis of probable alzheimer's disease over the last two decades: Ii. *Neurology*, 55:1863–1869, 2000.

[37] J. Pantel, K. Cretsinger, and H. Keefe. Hippocmapus tracing guidelines. Available at: http://www.psychiatry.uiowa.edu/ipl/pdf/hippocampus.pdf, 1998.

[38] RC Petersen, GE Smith, SC Waring, RJ Ivnik, E Kokmen, and EG Tangelos. Aging, memory, and mild cognitive impairment. *Int Psychogeriatr*, 9 (suppl. 1):65–69, 1997. Seminal paper on mild cognitive impairment.

[39] J.C. Pinheiro and D.M. Bates. *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing. Springer, 2000.

[40] A Pitiot, AW Toga, and PM. Thompson. Adaptive elastic segmentation of brain mri via shape-model-guided evolutionary programming. *IEEE Transactions on Medical Imaging*, 21(8):910–923, August 2002.

[41] SM Pizer, DS Fritsch, P Yushkevich, V Johnson, E Chaney, and G Gerig. Segmentation, registration, and measurement of shape variation via image object shape. *IEEE Transactions on Medical Imaging*, October 1999.

[42] J.P.W. Pluim, J.B.A. Maintz, and M.A. Viergever. Mutual information based registration of medical images: a survey. *IEEE Transactions on Medical Imaging*, In press.

[43] G Rizzo, P Scifo, M Gilardi, V Bettinardi, F Grassi, S Cerutti, and F. Fazio. Matching a computerized brain atlas to multimodal medical images. *NeuroImage*, 6:59–69, 1997.

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

[45] N Schuff, D Amend, F Ezekiel, SK Steinman, J Tanabe, D Norman, W Jagust, JH Kramer, JA Mastrianni, G Fein, and MW Weiner. Changes of hippocampal n-acetyl aspartate and volume in alzheimer's disease. a proton mr spectroscopic imaging and mri study. *Neurology*, 49:1513–1521, 1997.

[46] Dinggang Shen, Scott Moffat, Susan M. Resnick, , and Christos Davatzikos. Measuring size and shape of the hippocampus in mr images using a deformable shape model. *NeuroImage*, 15(2):422–434, February 2002.

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

[48] JP Thirion. Image matching as a diffusion process: an analogy with maxwell's demons. *Medical Image Analysis*, 2(3):243–260, 1998.

[49] P.M. Thompson, K.M. Hayashi, G. de Zubicaray, A.L. Janke, S.E. Rose, J. Semple, M.S. Hong, D. Herman, D. Gravano, D.M. Doddrell, and A.W. Toga. Mapping hippocampal and ventricular change in alzheimer's disease. *NeuroImage*, June 2004.

[50] N Tzourio-Mazoyer, B Landeau, D Papathanassiou, F Crivello, O Etard, and N Delcroix. Automated anatomical labelling of activations in spm using a macroscopic anatomical parcellation of the mni mri single subject brain. *NeuroImage*, 15:273–289, 2002.

[51] Simon K. Warfield, Andre Robatino, Joachim Dengler, Ferenc A. Jolesz, and Ron Kikinis. *Nonlinear Registration and Template Driven Segmentation*, chapter 4. Progressive Publishing Alternatives, 1998.

[52] Simon K. Warfield, Kelly H. Zou, and William M. Wells. Simultaneous truth and performance level estimation (staple): An algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7), July 2004.

[53] C Watson, F Andermann, P Gloor, M Jones-Gotman, T Peters, A Evans, A Olivier, D Melanson, and G Leroux. Anatomic basis of amygdaloid and hippocampal volume measurement by magnetic resonance imaging. *Neurology*, 42(9):1743–1750, 1992.

[54] J Webb, A Guimond, P Eldridge, D Chadwick, J Meunier, JP Thirion, and N Roberts. Automatic detection of hippocampal atrophy on magnetic resonance images. *Magnetic Resonance Imaging*, 17(8):1149–1161, October 1999.

[55] RP Woods, ST Grafton, CJ Holmes, SR Cherry, and JC Mazziotta. Automated image registration: I. general methods and intrasubject, intramodality validation. *Journal of Computer Assisted Tomography*, 22:139–152, 1998.

[56] Terry S. Yoo, editor. *Insight into Images*. Insight Software Consortium, 2004.