ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 27 (2005) 979 – 990

Atlas-based hippocampus segmentation in Alzheimer's disease and mild cognitive impairment

Owen T. Carmichael,[a,*] Howard A. Aizenstein,[b] Simon W. Davis,[a] James T. Becker,[b,c,f] Paul M. Thompson,[d] Carolyn Cidis Meltzer,[a] and Yanxi Liu[e]

[a]*Radiology Department, University of Pittsburgh, B-938 PUH, 200 Lothrop Street, Pittsburgh, PA 15213, USA*
[b]*Psychiatry Department, University of Pittsburgh, Pittsburgh, PA 15213, USA*
[c]*Neurology Department, University of Pittsburgh, Pittsburgh, PA 15213, USA*
[d]*Neurology Department, University of California, Los Angeles, Los Angeles, CA 90095-1769, USA*
[e]*The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[f]*Psychology Department, University of Pittsburgh, Pittsburgh, PA 15213, USA*

Received 26 January 2005; revised 13 March 2005; accepted 3 May 2005
Available online 28 June 2005

This study assesses the performance of public-domain automated methodologies for MRI-based segmentation of the hippocampus in elderly subjects with Alzheimer's disease (AD) and mild cognitive impairment (MCI). 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 ("cohort atlases"). 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 spatial overlap between automated segmentations and expert manual segmentations. Registration methods that produced higher degrees of geometric deformation produced automated segmentations with higher agreement with manual segmentations. Side of the brain, presence of AD, choice of reference image, and manual tracing protocol were also significant factors contributing to automated segmentation performance. Fully automated techniques can be competitive with human raters on this difficult segmentation task, but a rigorous statistical analysis shows that a variety of methodological factors must be carefully considered to insure that automated methods perform well in practice. The use of fully deformable registration methods, cohort atlases, and user-defined manual tracings are recommended for highest performance in fully automated hippocampus segmentation.

Keywords: Hippocampus segmentation; Alzheimer's disease; Mild cognitive impairment

Introduction

Hippocampal atrophy has been proposed as a clinical marker for early AD because it is known to occur early in the course of the disease on a spatial scale large enough to be detectable with structural MR images (Bobinski et al., 1996; Kordower et al., 2001). Visual qualitative atrophy assessment (e.g., de Leon et al., 1993) has been hindered by the relative subtlety of atrophy early in the course of AD (Frisoni, 2001). However, the development of reliable repeatable protocols for human raters to trace the hippocampus (e.g., Jack et al., 1995) has led to the possibility of precise quantitation of AD-related atrophy (Chetelat and Baron, 2003), hippocampus-level quantification of activation in co-registered structural–functional images (Dickerson et al., 2004), and quantification of other hippocampal characteristics such as bilateral symmetry (Bigler et al., 2002). Furthermore, tracing protocols have enabled the study of hippocampal morphometrics in subjects with mild cognitive impairment (MCI), a high-AD-risk clinical condition marked by minor deficits in one or more cognitive domains (Convit et al., 1997; Petersen et al., 1997).

However, large-scale studies of AD-related hippocampal atrophy are often impractical because manual segmentations are labor-intensive and require training to insure high repeatability between raters. Typical hippocampi take between 30 min and 2 h to trace by hand; the tedious labor quickly causes fatigue. Semi-automated segmentation methods reduce manual labor by having the user identify a sparse set of image landmarks that constrain a subsequent automated segmentation process (Christensen et al., 1997; Freeborough et al., 1997; Shen et al., 2002). However, we focus on fully automated atlas-based 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.

* Corresponding author. Fax: +1 412 647 0700.
E-mail address: otc@andrew.cmu.edu (O.T. Carmichael).
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2005.05.005

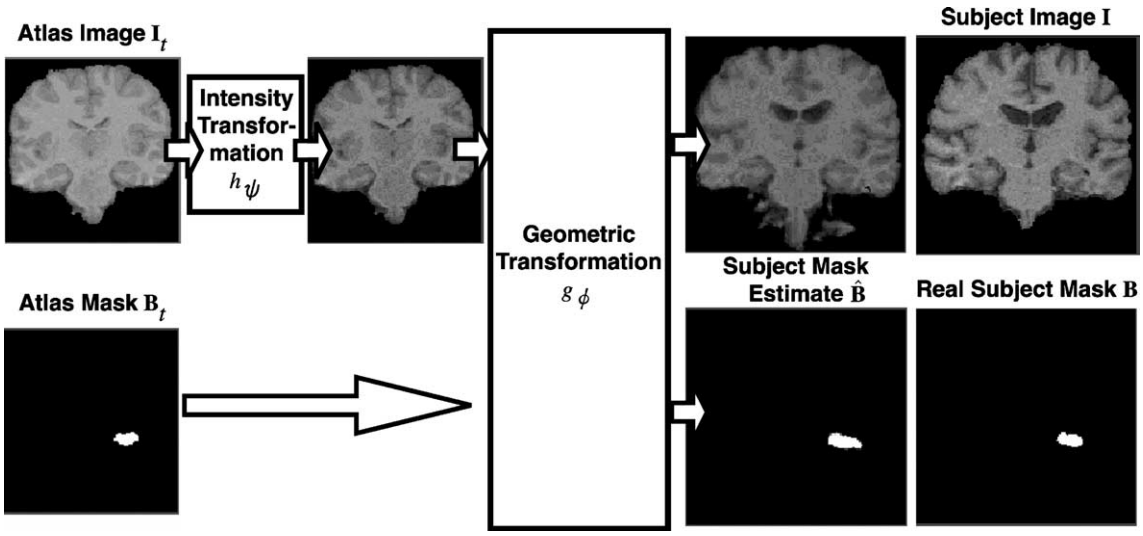

Fig. 1. Schematic view of atlas-based segmentation. An intensity transformation and geometric transformation are estimated to register the atlas image to the subject image; the geometric transformation is applied to the atlas mask in order to estimate the subject mask.

Atlas-based segmentation coregisters a subject image and a special reference image called the *atlas image* on which structures of interest have been manually traced (see Fig. 1). The resulting spatial transformation maps the coordinates of the structures from the coordinate space of the atlas image to that of the subject image. Since this approach is posed in terms of image-to-image registration, atlas-based techniques take advantage of methodological advances in registration that are driven by a wide range of application areas such as visualization, image-guided surgery, and voxel-based morphometry. Furthermore, atlas-based approaches are among the easiest to implement since they only require the user to align the atlas and subject images.

The purpose of this study was to systematically compare the performance of several competing public-domain methodologies for atlas-based segmentation of AD-atrophied hippocampi. We validated several widely disseminated automated image registration methods (Friston et al., 1995; Jenkinson et al., 2002; Woods et al., 1998); in contrast, previous studies on atlas-based elderly hippocampus segmentation used a single, recently developed, cutting-edge registration algorithm that lacked a widely disseminated, standard software implementation (e.g., Crum and Scahill, 2001). Furthermore, we examined the use of two widely disseminated atlas images (Kikinis et al., 1996; Tzourio-Mazoyer et al., 2002), as well as individual manually traced subject images (as in, e.g., Webb et al., 1999) to serve as the reference or *cohort atlas* image. Finally, we examined the impact of varying manual tracing protocols on atlas-based segmentation performance.

Methods

Subject data

We gathered 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, had the following parameters optimized for maximal contrast among gray matter, white matter, and CSF (TE = 5, TR = 25, flip angle = 40°, NEX = 1, slice thickness = 1.5 mm/0 mm 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 (Lopez et al., 2000a,b) for evaluation procedure). A consensus meeting of neuroradiologists, psychiatrists, neurologists, and psychologists diagnosed each subject into MCI (Petersen et al., 1997), AD, or control categories.

Skulls were stripped from all images using the Brain Extraction Tool (BET) (Smith, 2002), and the images were cropped to remove all-zero slices using the crop tool provided with AIR 2.0 (Woods et al., 1998).

Registration methods

We compared the performance of software modules in AIR, SPM, FLIRT, and Chen's method (Woods et al., 1998; Friston et al., 1995; Jenkinson et al., 2002; Chen, 1999) as registration substrates for atlas-based segmentation of elderly hippocampi.[1] While several algorithmic details vary between these registration techniques, they are chiefly distinguished from each other in terms of their *geometric transformation model*—that is, the mathematical equation that maps image coordinates between the atlas image and subject image. We partitioned the geometric transformation models into three categories in terms of the degree to which they allow the atlas image to spatially deform when it is aligned to the subject image. *Affine* methods apply the same linear transformation to all voxels in the entire atlas image; *semi-deformable* mappings deform the atlas image in a spatially smooth gradual way to align it to the subject image; and *fully deformable* methods produce image-to-image mappings that are essentially unconstrained spatially (see Fig. 2 for an illustration

[1] While we use the original implementation of Chen's method, its registration modules are highly similar to the FEM-based and Demons registration modules of the freely available ITK software package (Yoo, 2004).

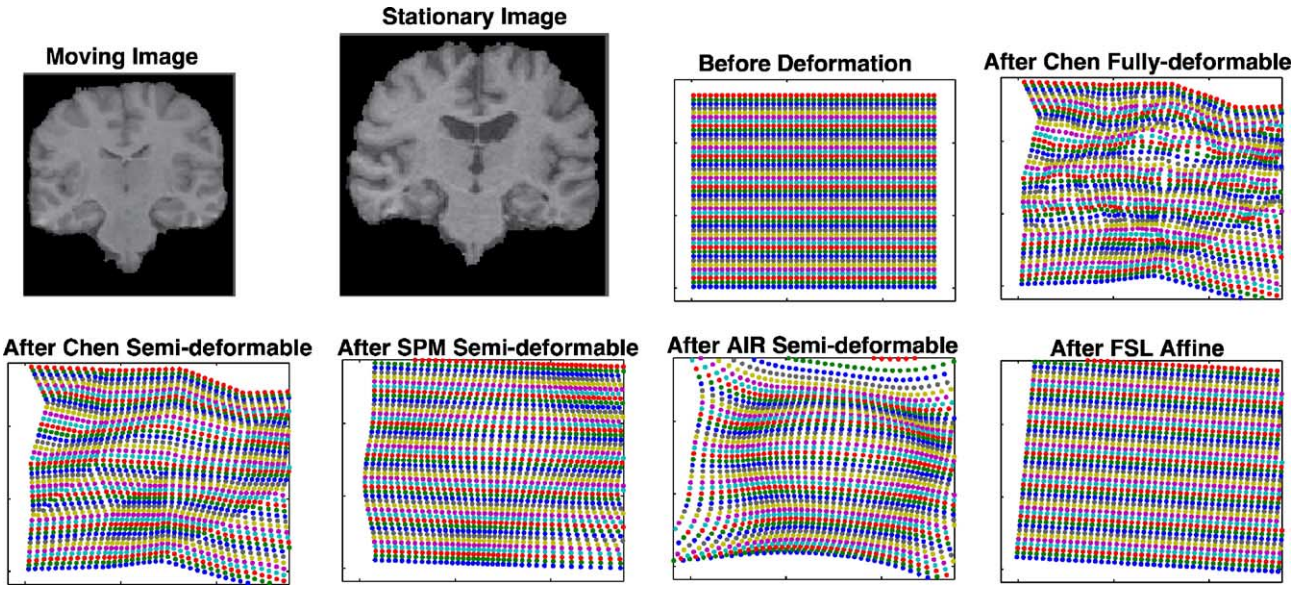

Fig. 2. Example image deformations produced by fully deformable, semi-deformable, and affine registration techniques. The moving image is registered to the stationary image using each of the 7 algorithms we analyze. The colored dots show the geometric positions of voxels in the shown slice of the moving 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 Carmichael et al., 2004 for a detailed mathematical explanation). We considered one fully deformable, three semi-deformable, and three affine registration methods. The *AIR affine* (Woods et al., 1998), *SPM affine* (Friston et al., 1995), and *FLIRT affine* (Jenkinson et al., 2002) methods estimate an affine transformation between images. The *AIR semi-deformable* method uses the transformation output by the AIR affine method as a starting point for estimation of a spatially smooth deformation based on a polynomial transformation model; the *SPM semi-deformable method* uses the transformation output by the SPM affine method as the starting point for estimation of a smooth deformation based on a discrete cosine transform (DCT) transformation model; the *Chen semi-deformable* method estimates a piecewise-linear transformation (Chen, 1999). Finally, the *Chen fully deformable* method takes the output of the Chen semi-deformable method as a starting point for estimation of an unconstrained voxel-by-voxel deformation.

Manual segmentations

We evaluated 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. (1992), Schuff et al. (1997), and Pantel et al. (1998). 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. Rater R1 traced hippocampi on all 54 subject images; additionally, 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–manual segmentation agreement to inter-rater agreement. All manual segmentations were digitized into binary volumes for analysis.

Cohort atlases

In the cohort atlas scenario, we selected an image from a subject population (AD, MCI, or control), manually traced left and right hippocampi on it, and treated it as a reference image that all other images in the subject population were registered to during atlas-based segmentation. We refer to the selected subject image as a cohort "atlas" image to emphasize its role as a reference image. Cohort atlas images were selected at random from the subject population, however, we note that a variety of more complex strategies for cohort atlas image selection are possible (Rohlfing et al., 2004). For each image in each subject population, we considered a hypothetical situation in which that image is selected as the cohort atlas; all other images in the population were registered to the cohort atlas image, and hippocampus segmentation results were evaluated. In other words, for a population of k images, we considered k different possible cohort atlases, which we registered to all $k - 1$ other images in the population for a total of $k - 1$ trials per registration method.

Standard atlases and atlas tracers

In the standard atlas scenario, we began with an atlas image and manual hippocampus tracings, or *manual atlas tracings*, provided by an atlas institution (Harvard or MNI). We registered the atlas image to the subject image, and we used the resulting transformation to transfer a manual tracing of the hippocampus from the atlas image to the subject image. This automated segmentation was evaluated by comparing it to an independent *manual subject tracing*—a manual tracing of the hippocampi on the subject image performed by rater R1. However, we recognized that the manual tracing protocol used by R1 may differ from that used by human tracers at MNI and Harvard and that our evaluation risked confounding two factors that could have caused discrepancies between the automated segmentation and manual subject tracing: differences in hippocampus delineation between automated and manual techniques and discrep-

ancies in hippocampus boundary conventions between manual atlas and subject tracings. For this reason, rater R1 generated manual atlas tracings by tracing left and right hippocampi on the Harvard and MNI atlas images using the same manual tracing protocol used for tracing on the subject images. Experiments analyzed the effects of choice of atlas (MNI vs. Harvard) and manual atlas tracings (performed by R1 vs. performed by the atlas institution) on manual–automated segmentation agreement.

Cohort atlas images reflect possibly anomalous characteristics of a particular scan and subject, and their use is inherently more labor-intensive than standard atlases since they require the user to hand-label the structure of interest on the cohort atlas image. However, cohort-atlas-based segmentation has potential advantages over the more conventional standard-atlas-based approach. If the population of subject 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 on the cohort atlas image insures the investigator that anatomical boundaries reflect his or her conventions.

Manual–automated and manual–manual agreement

Performance of automated segmentation algorithms was measured in terms of *manual–automated agreement*, that is, agreement between automated segmentations and manual tracings performed by an expert rater. We compared manual–automated agreement to *manual–manual agreement*, or the agreement between manual tracings performed by pairs of expert human raters. In doing so, we assessed 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 R1 traced hippocampi on the full set of 54 subject images, we measured manual–automated agreement in terms of agreement between R1-rated manual tracings and the Chen fully deformable automated technique. Manual–manual agreement was measured in terms of pairwise agreement between manual tracings by R1 and R2, R1 and R3, and R2 and R3. Manual–automated agreement for each subject image was summarized in terms of the average manual–automated agreement between its R1 segmentation and the automated segmentations from all cohort atlas images in its disease category. Experiments analyzed differences between manual–manual agreement and manual–automated agreement on the 6 multiply manually traced hippocampi. Note that this approach differs from the more common approach of measuring agreement between pairs of manual and/or automated segmentations 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. We note that other approaches, based on estimating automated segmentation performance and a single estimate of the true, underlying structure mask, are also available (Warfield et al., 2004).

Performance measure: overlap ratio

We evaluated the agreement between an automated hippocampus segmentation estimate and a manual segmentation using a numerical criterion that measured the degree to which they overlap. We represented the automated segmentation $\hat{\mathbf{B}}$ and its corresponding manual segmentation $\mathbf{B}$ as binary 3D volumes in which voxels labeled as hippocampus had a value of 1. Let V_{BOTH} be the set of voxels labeled as hippocampus by both $\hat{\mathbf{B}}$ and $\mathbf{B}$; set $V_{\hat{\mathbf{B}}}$ has voxels labeled as hippocampus by $\hat{\mathbf{B}}$ but not $\mathbf{B}$; and set $V_{\mathbf{B}}$ consists of voxels labeled as hippocampus by $\mathbf{B}$ but not $\hat{\mathbf{B}}$ (sets V_{BOTH}, $V_{\hat{\mathbf{B}}}$, and $V_{\mathbf{B}}$) are labeled in red, dark gray, and light gray in Fig. 3d). The *overlap ratio* measures the degree of overlap between the automated and manual segmentations, specifically:

$$or\left(\mathbf{B},\hat{\mathbf{B}}\right)=\frac{|V_{\mathrm{BOTH}}|}{|V_{\mathrm{BOTH}}|+|V_{\hat{\mathbf{B}}}|+|V_{\mathbf{B}}|}$$

In other words, the overlap ratio measures the percentage of the combined volumes of $\hat{\mathbf{B}}$ and $\mathbf{B}$ that are both labeled as hippocampus. When $\hat{\mathbf{B}}$ and $\mathbf{B}$ overlap perfectly, $or(\hat{\mathbf{B}}, \mathbf{B}) = 1$; when the masks do not overlap at all, $or(\hat{\mathbf{B}}, \mathbf{B}) = 0$. We note that several authors have quantified manual–automated segmentation agreement using criteria similar to the overlap ratio (Dawant et al., 1999; Kelemen et al., 1999; Klemencic et al., 2001; Shen et al., 2002).

Overlap ratio was computed over the entire hippocampus. Furthermore, to characterize automated segmentations in terms of hippocampal sub-regions, we divided the hippocampus into sections and computed performance measures over the voxels in each section. Consider a bounding box ($x_{\min}$, $x_{\max}$, $y_{\min}$, $y_{\max}$, $z_{\min}$, $z_{\max}$) around all the hippocampus voxels in $\hat{\mathbf{B}}$ and $\mathbf{B}$ (i.e., the x coordinates of all voxels in $V_{\mathrm{BOTH}} \cup V_{\hat{\mathbf{B}}} \cup V_{\mathbf{B}}$ are between $x_{\min}$ and $x_{\max}$, etc.). For each of the three cardinal directions, we partitioned the estimated and ground-truth hippocampi into k sections along that direction and computed overlap ratios in each of the sections. That is, for all i from 1 to k, we computed $or(\mathbf{B}_i^x, \hat{\mathbf{B}}_i^x)$, 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 $\mathbf{B}_i^x(x,y,z) = 0$ and for all other voxels. Similarly, we compute $(\mathbf{B}_i^y, \hat{\mathbf{B}}_i^y)$ and $or(\mathbf{B}_i^z, \hat{\mathbf{B}}_i^z)$ for all i from 1 to k. See Fig. 3e for an illustration. Fig. 4 suggests that, since the hippocampi all have similar gross 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, the sections to the left correspond to posterior regions, and sections to the right correspond to anterior regions.

Statistical analysis: mixed-effects models

We analyzed the effects of registration method, side of the brain, disease state, manual tracing protocol, and choice of atlas on overlap ratio through mixed-effects statistical models (Pinheiro and Bates, 2000) that properly accounted for fixed effects, random effects, and grouping in our data. The fixed effects, including disease state, side of the brain, and registration method, were 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, were modeled as variance components. Each level of each fixed effect was assigned a coefficient representing the offset it produced from the baseline value. The overall significance of each fixed effect was evaluated through omnibus F tests. Furthermore, we analyzed differences between factor levels–for example, between control, MCI, and AD subjects—by using focused F tests to check for significant differences between their coefficients. Effect size for focused F

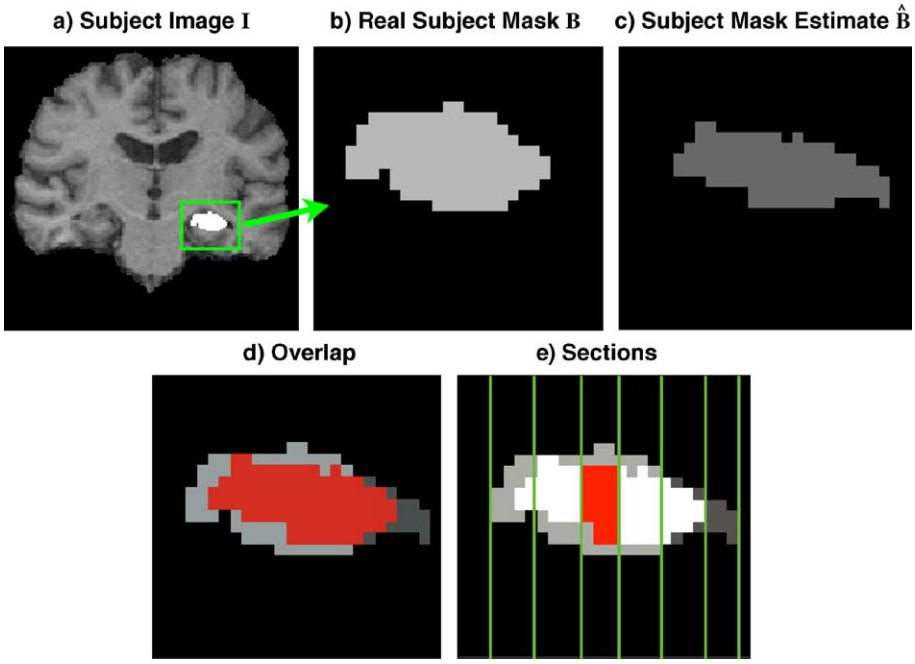

Fig. 3. Evaluating consistency between masks using overall and sectional overlap. A ground-truth subject mask and estimated subject mask are shown in light and dark gray. (d) Voxels in red overlap between the ground-truth and the estimate. Overlap ratio measures the ratio between the volume of the red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

tests was quantified by the contrast correlation r_{contrast} (Rosenthal et al., 2000), which generalizes standard 2-group correlational effect size measures while properly accounting for degrees of freedom. In our analysis, between-group differences refer to differences in model coefficients between two factor levels. Mixed-effects models properly account for the random sampling of subject images and cohort atlas images from overall AD, MCI, and control populations and properly account for repeated measures. All statistics were performed using R version 1.9.1. Mixed-effects models were fit using maximum likelihood estimation in the nlme package. In order to give multiple views of the complex ways in which overlap ratio varied with respect to fixed effects, we report significance values and effect sizes for between-group tests, as well as box-and-whisker plots that show the median, quartiles, and extreme values of overlap ratio within groups.

Experiments evaluate the degree to which segmentation results varied with respect to disease state, registration algorithm, atlases, manual tracings, and side of the brain. At the core of the 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

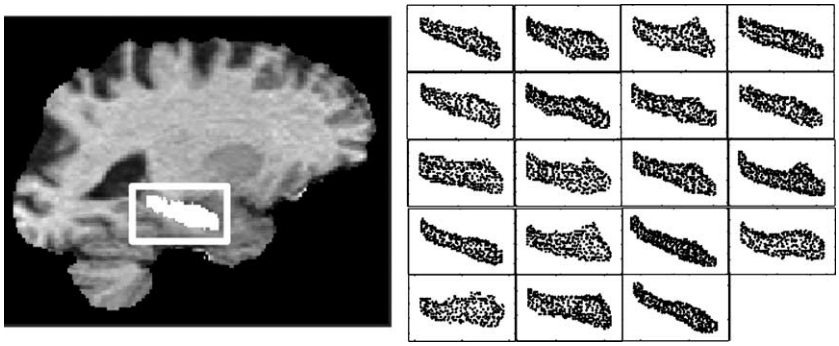

Fig. 4. Points on the left hippocampus in all 19 MCI subjects are shown projected onto the *xz* plane of the image. Note that all the hippocampi share the same rough initial orientation in this plane.

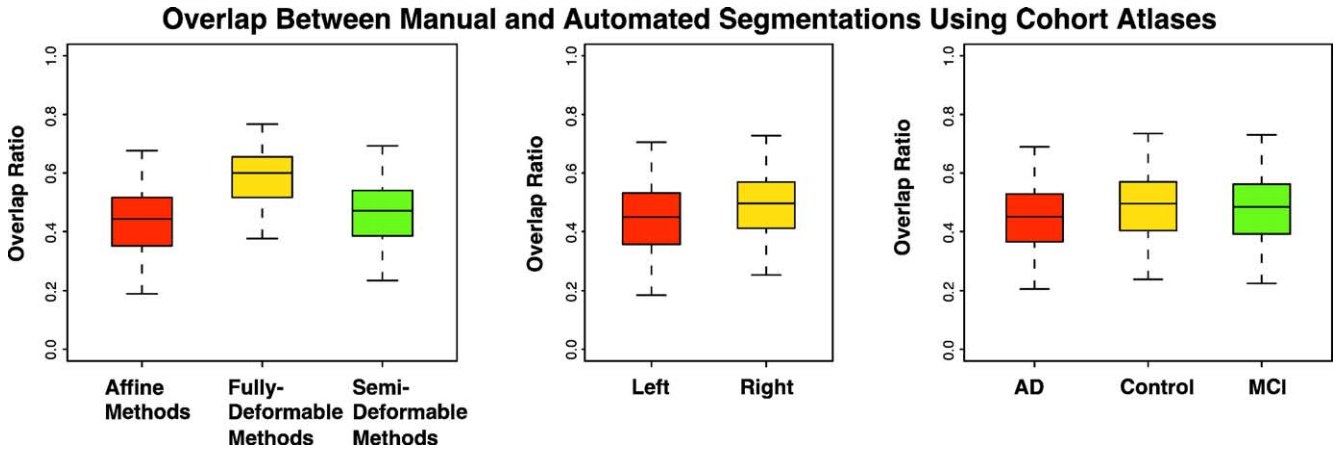

Fig. 5. Overlap ratio as a function of disease state, registration method category, and side of the brain for the 54 images using cohort atlases.

manual tracings supplied with the atlas. For each disease state, and for each registration algorithm, we ran one trial for each possible cohort atlas image and subject image within the disease group.

Results

Cohort atlases

For cohort-atlas-based segmentation, we fit a mixed-effects model 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 ($P < 0.0001$, $P = 0.0192$, $P < 0.0001$).

Box plots showing how overlap ratio varies with disease state, side of the brain, registration method, and registration method category are shown in Fig. 5. Effect sizes and P values are shown in Table 1. Overlap ratio was significantly lower in AD compared to MCI and control groups, although no significant difference in overlap ratio was seen between MCI and control groups. No significant difference existed between the FLIRT affine and AIR affine methods. For all other pairs of methods, significant (but in many cases slight) differences in overlap ratio existed. 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.

Comparing fully deformable, semi-deformable, and affine methods

We grouped the registration methods into fully deformable, semi-deformable, and affine categories 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

Table 1
P values and the contrast correlation $r_{contrast}$ (Rosenthal et al., 2000) are shown for F tests between pairs of registration methods, disease states, and method categories in the mixed-effects model for cohort-atlas-based segmentation

P, $r_{contrast}$	Chen semi	AIR semi	SPM semi	AIR affine	SPM affine	FLIRT affine
Chen fully	<0.0001, 0.401***	<0.0001, 0.296***	<0.0001, 0.445***	<0.0001, 0.477***	<0.0001, 0.431***	<0.0001, 0.470***
Chen semi		<0.0001, 0.127***	<0.0001, 0.060***	<0.0001, 0.104***	<0.0001, 0.040***	<0.0001, 0.095***
AIR semi	<0.0001, 0.127***		<0.0001, 0.184***	<0.0001, 0.227***	<0.0001, 0.165***	<0.0001, 0.218***
SPM semi	<0.0001, 0.060***	<0.0001, 0.184***		<0.0001, 0.045***	0.029, 0.020*	<0.0001, 0.035***
AIR affine	<0.0001, 0.104***	<0.0001, 0.227***	<0.0001, 0.045***		<0.0001, 0.065***	0.286, 0.010
SPM affine	<0.0001, 0.040***	<0.0001, 0.165***	0.029, 0.020*	<0.0001, 0.065***		<0.0001, 0.055***
FLIRT affine	<0.0001, 0.095***	<0.0001, 0.218***	<0.0001, 0.035***	0.286, 0.010	<0.0001, 0.055***	

P, $r_{contrast}$	MCI	AD
Control	0.647, 0.064	0.011, 0.345*
MCI		0.024, 0.310*
AD	0.024, 0.310*	

P, $r_{contrast}$	Semi-Deformable	Affine
Fully Deformable	<0.0001, 0.450***	<0.0001, 0.530***
Semi-Deformable		<0.0001, 0.170***
Affine	<0.0001, 0.170***	

Between left and right sides of the brain, P and $r_{contrast}$ are <0.0001 and 0.277 respectively.
P values and effect sizes for differences in manual–automated overlap using cohort atlases.
* $P < 0.05$.
*** $P < 0.001$.

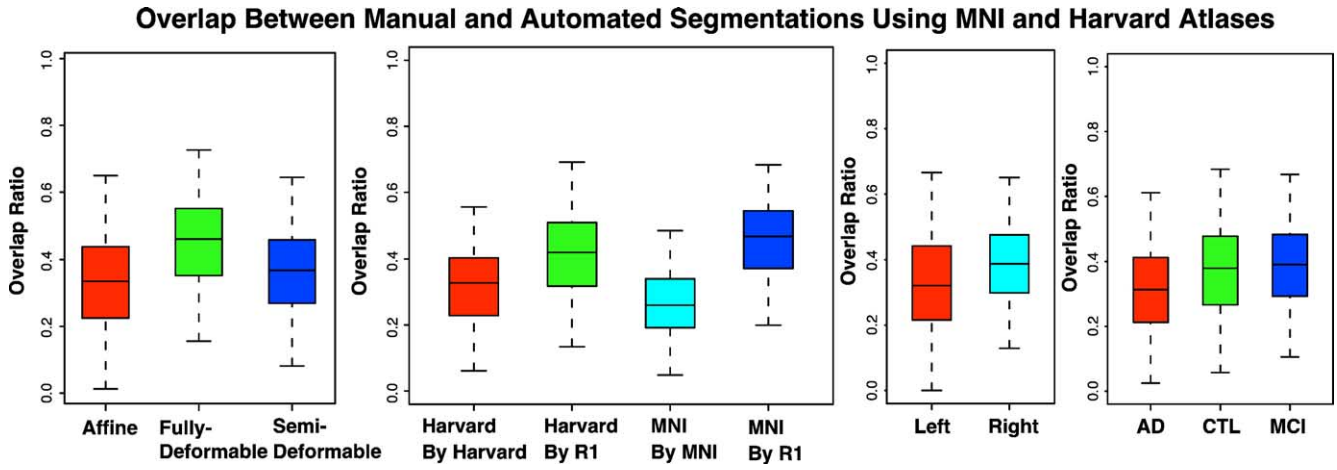

Fig. 6. Overlap ratio as a function of disease state, registration method, manual tracing, and side of the brain for the 54 images using standard atlases.

effects (see Fig. 5 and Table 1). Fully deformable methods had significantly higher overlap ratio than semi-deformable and affine methods. In turn, semi-deformable methods had significantly higher overlap ratio than affine methods, although the effect size was not as pronounced as in the comparison between fully deformable and semi-deformable categories.

Standard atlases and manual atlas tracings

For standard-atlas-based segmentation, we fit a mixed-effects model in which the fixed effects were the standard atlas (Harvard or MNI), the source of the manual atlas tracings (R1 or Harvard/MNI), side of the brain, disease state, and registration method; subject identity was a random effect; and the performance measures were dependent variables. Fig. 6 and Table 2 present the overlap ratio as a function of atlas image and manual atlas tracing, registration method, side of the brain, and disease state. Results based on R1-traced atlas tracings are referred to as "Harvard By R1" and "MNI By R1"; results based on manual atlas tracings 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 manual atlas tracings than hippocampi traced by the atlas institution and was significantly higher for right sides of the brain compared to left. No significant difference in overlap ratio was seen between the MNI and Harvard atlases. Overlap ratio was significantly lower for AD subjects than MCI subjects and controls, but no significant difference was seen between the MCI and control groups. The

Table 2
P values and the contrast correlation $r_{contrast}$ for *F* tests between factor levels in the mixed-effects model for standard-atlas-based segmentation

P, $r_{contrast}$	Chen semi	AIR semi	SPM semi	AIR affine	SPM affine	FLIRT affine
Chen fully	<0.0001, 0.235***	<0.0001, 0.204***	<0.0001, 0.374***	<0.0001, 0.306***	<0.0001, 0.341***	<0.0001, 0.395***
Chen semi		0.072, 0.033	<0.0001, 0.159***	<0.0001, 0.306***	<0.0001, 0.120***	<0.0001, 0.184***
AIR semi	0.072, 0.033		<0.0001, 0.191***	<0.0001, 0.112***	<0.0001, 0.152***	<0.0001, 0.215***
SPM semi	<0.0001, 0.159***	<0.0001, 0.191***		<0.0001, 0.0817***	0.027, 0.041*	0.163, 0.026
AIR affine	<0.0001, 0.306***	<0.0001, 0.112***	<0.0001, 0.0817***		0.024, 0.041*	<0.0001, 0.107***
SPM affine	<0.0001, 0.120***	<0.0001, 0.152***	0.027, 0.041*	<0.0001, 0.065***		<0.0003, 0.066***
FLIRT affine	<0.0001, 0.184***	<0.0001, 0.215***	0.163, 0.026	<0.0001, 0.107	0.0003, 0.066***	

P, $r_{contrast}$	MCI	AD
Control	0.665, 0.061	0.020, 0.319*
MCI		0.004, 0.390**
AD	0.004, 0.390**	

	P, $r_{contrast}$
R1 vs. other tracers	<0.0001, 0.642***
Left vs. right	<0.0001, 0.331***
Harvard vs. MNI	0.900, 0.0023

P, $r_{contrast}$	Semi-def.	Affine
Fully def.	<0.0001, 0.326***	<0.0001, 0.410***
Semi-def.		<0.0001, 0.146***
Affine	<0.0001, 0.146***	

P values and effect sizes for differences in manual–automated overlap using standard atlases.
* $P < 0.05$.
** $P < 0.01$.
*** $P < 0.001$.

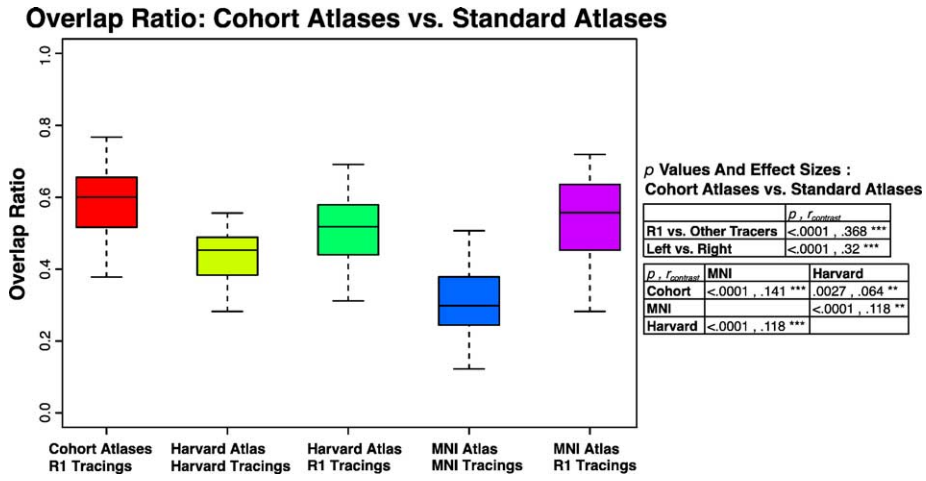

Fig. 7. Left: Overlap ratio for cohort-atlas-based and standard-atlas-based segmentation using Chen's fully deformable registration method. Right: *P* values and the contrast correlation r_{contrast} for *F* tests between factor levels in the mixed-effects model.

registration methods, ranked in decreasing order of overlap ratio, 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 nor was the difference in overlap ratio between the Chen semi-deformable method and AIR semi-deformable method. Differences in overlap ratio between all other pairs of methods had statistically significant *P* values, although in some cases the effect sizes were not large.

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 above. We fit a mixed-effects model in which the atlas (MNI, Harvard, or cohort atlas), human tracer (R1 or the atlas institution), side of the brain, and disease state were fixed effects, subject identity was a random effect, and the dependent variable was the overlap ratio. Fig. 7 plots the overlap ratio for the standard atlases and cohort atlases in this model, along with *P* values and effect sizes. The mean overlap ratio was significantly higher for cohort-atlas-based segmentation than standard-atlas-based-segmentation using manual atlas tracings by R1 along with the MNI or Harvard atlas images. Performance measures for standard atlases using manual atlas tracings from the atlas institution were significantly worse in each case.

Manual–automated agreement and manual–manual agreement

For the six multiply manually traced subjects, we fit a mixed-effect model with overlap ratio as the dependent variable, the type of agreement (manual–manual or manual–automated), and side of the brain as fixed effects, and subject identity as a random effect. Manual–manual agreement was not significantly higher than manual–automated agreement in terms of overlap ratio, although we saw a trend toward slightly higher manual–manual agreement ($P = 0.0916$, $r_{\text{contrast}} = 0.264$). Box plots comparing the distribution of overlap ratio for manual–manual and manual–automated agreement are shown in Fig. 8.

Sectional results

Fig. 9 shows a representative plot of automated–manual mean overlap ratios and manual–manual mean overlap ratios for

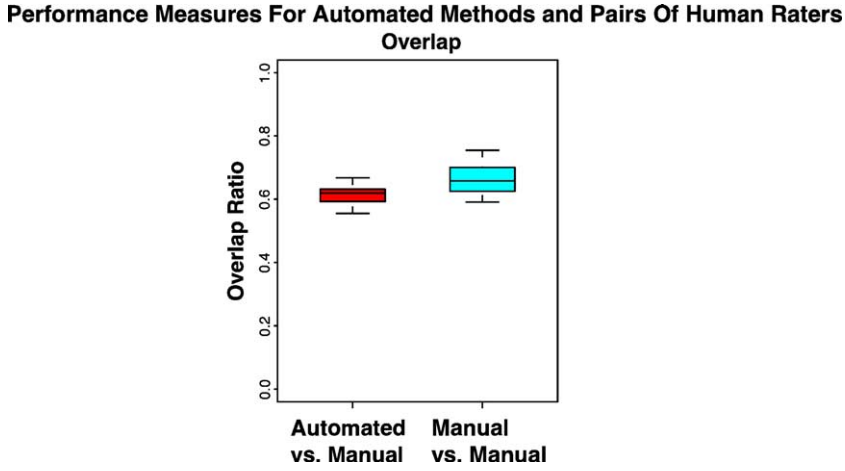

Fig. 8. Overlap ratio between manual and automated segmentations (automatic vs. manual) and between pairs of manual segmentations (manual vs. manual).

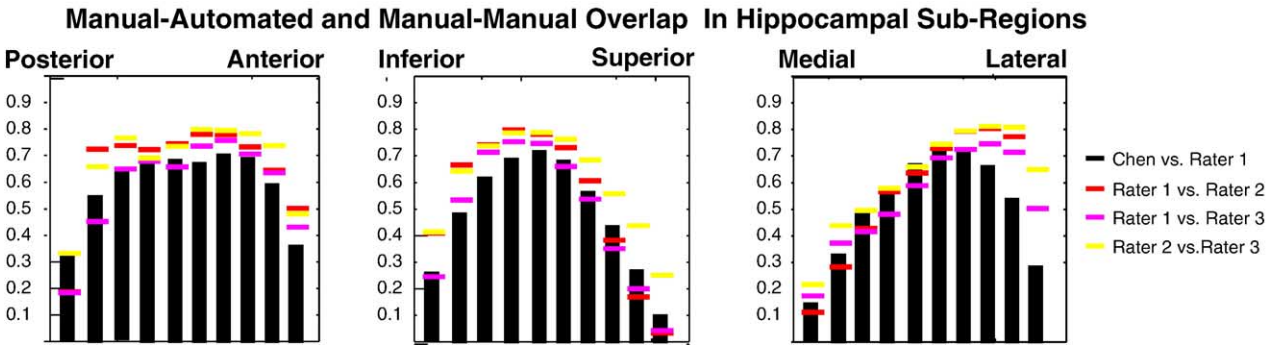

Fig. 9. Overlap ratio delineated along posterior–anterior line (left), inferior–superior line (middle), and medial–lateral line (right) for the Chen fully automated registration method on the right hippocampus in MCI images. Similar patterns of overlap ratio distribution are seen for other registration methods, other disease states, and the left hippocampus. See text for details.

hippocampal sections taken along the three cardinal directions of our data set. Results are shown for cohort-atlas-based segmentation using the Chen fully deformable registration method on the right hippocampus in MCI images; however, similar distributions of overlap are seen for both sides of the brain, all registration methods, and all disease states (see [3] for detailed plots). The three cardinal directions correspond roughly to the posterior–anterior, medial–lateral, and superior–anterior hippocampal axes, respectively (see Fig. 4). The hippocampal sections most responsible for manual–automated disagreement were 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 was generally higher toward the lateral extent of the hippocampus and lower toward the medial extent. Furthermore, with the exception of the most extreme sections, mean overlap ratio was relatively constant with respect to anterior–posterior position. Finally, moving from the superior to inferior extent, mean overlap ratio increased steadily, reached a peak at the central sections, and decreased toward the inferior end. These patterns of manual–automated overlap across sub-regions were similar to patterns of manual–manual overlap on the 6 selected images, although the human raters were relatively more consistent at the lateral extent.

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 for the first factor level compared to the second.

Fully deformable > semi-deformable ≥ affine

Our results confirm the intuition that methods making use of more highly deformable geometric transformation models tended to fit the complex shape of the hippocampus more accurately than less-deformable geometric models. This agrees with earlier results that demonstrated that atlas-based hippocampus segmentation based on other highly deformable registration methods can produce hippocampal volumes consistent with expected disease-related atrophy effects (see, e.g., (Crum and Scahill, 2001; Fischl et al., 2002; Hogan et al., 2004). 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. In a related technical report, a statistical analysis of the severity of segmentation errors shows a similar relationship between the performance characteristics of the three registration categories (Carmichael et al., 2004).

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 Figs. 9 and 8), 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 ratio. These results, together with results from a related study of the severity of automated segmentation errors (Carmichael et al., 2004), suggest that automated methods may be competitive for elderly hippocampus segmentation applications, especially those that can tolerate minor errors in spatial localization. These results extend previous findings that atlas-based techniques can be competitive with manual tracing for other subject groups and brain structures (see, e.g., Chard et al., 2002; Collins et al., 1995; Leemput et al., 1999; Warfield et al., 1998). Furthermore, the automated results present a very promising starting point for further automated refinement by more complex shape-model-based segmentation techniques (for example, Kelemen et al., 1999; Pitiot et al., 2002; Pizer et al., 1999).

MCI ≈ controls > AD

Overall performance measures were significantly lower among AD subjects than MCI or control subjects. One possible explanation for these results is that the degenerative processes of AD made 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 was 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 were similar could suggest that their image characteristics do not differ so significantly that they affected registration.

To further investigate how performance differences between disease groups were modulated by other algorithmic factors, we fit mixed-effects models similar to those described above, but with additional terms to model interactions between disease states and concurrent algorithmic factors. Specifically, for cohort-atlas-based segmentation, there were fixed effect terms in the model for disease state, side of the brain, registration method, and the interaction between disease state and registration method. For standard-atlas-based segmentation, we included fixed effect terms for the standard atlas, source of the manual atlas tracing, side of the brain, disease state, registration method, interaction between disease state and registration method, and interaction between disease state and standard atlas. For comparing standard atlases to cohort atlases using fully deformable registration, fixed effects were the atlas, human tracer, side of the brain, disease state, and interaction between disease state and atlas. Cohort-atlas-based segmentation performance was significantly higher for control subjects with the SPM affine ($P = .0232$, $r_{contrast} = .0206$), AIR semi-deformable ($P = .0035$, $r_{contrast} = .0265$), and SPM semi-deformable methods ($P < 0.0001$, $r_{contrast} = .0524$); standard-atlas-based performance was higher in control subjects with the AIR semi-deformable ($P = .0207$, $r_{contrast} = .0426$) and SPM semi-deformable methods ($P = .0128$, $r_{contrast} = .0458$) and lower in MCI subjects with the Chen semi-deformable method ($P = .0463$, $r_{contrast} = .0367$); and no interaction terms were significant in the model comparing cohort-atlas-based to standard-atlas-based segmentation using the Chen fully deformable method. While the effect size is relatively low for each interaction term, these results suggest that performance differences between AD, MCI, and control groups are attributable more to AIR and SPM registration methods than other registration methods or atlas choices.

Right > left

In terms of automated segmentation performance, a striking bilateral asymmetry was 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. (2002). However, a mixed-effects model fit to solely manual–manual agreement data did not show a statistically significant bilateral asymmetry in manual–manual overlap ratio ($P = .12$). Our initial calculations of hippocampal volumes did not show a significant volume asymmetry, echoing the findings of Bigler et al. (2002). Further investigation is needed to explain this bilateral effect.

Cohort-atlas-based ≥ 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 young healthy atlas-image brains and elderly diseased subject-image brains can negatively impact performance of standard-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 ≈ anterior, lateral > medial, center > periphery

Segmentation errors were evenly distributed between posterior and anterior regions of the hippocampus, were more concentrated in the medial regions than the lateral regions, and were 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. 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 did 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. (2004), 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 less sensitive in anterior sub-regions where radial distances are relatively large.

Manual tracing protocols add significant variability

Geuze et al. recently described a dizzying array of existing methods for manually tracing the hippocampus in MR (Geuze et al., 2004). Our results (see Fig. 6) 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 tracings and atlas-based automated methods. 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 recommend that researchers using standard atlas images for atlas-based segmentation should examine the manual tracings 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 image or a randomly selected subject image leads to automated segmentations whose agreement with expert manual segmentations is competitive with manual–manual agreement.

Conclusion

Atlas-based segmentation is a simple automated method for structure segmentation that can use public-domain tools to produce reasonable structure delineations in images of elderly controls and

subjects with MCI and AD. While additional work may be needed to make these automated techniques truly competitive with expert human raters, their performance may be acceptable for image processing 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.

Acknowledgments

This work was supported by NIH grants NS07391, MH064625, AG05133, DA01590001, MH01077, EB001561, RR019771, RR021813, and AG016570.

References

Bigler, E.D., Tate, D.F., Miller, M.J., Rice, S.A., Hessel, C.D., Earl, H.D., Tschanz, J.T., Plassman, B., Welsh-Bohmer, K.A., 2002. Dementia, asymmetry of temporal lobe structures, and apolipoprotein e genotype: relationships to cerebral atrophy and neuropsychological impairment. J. Int. Neuropsychol. Soc. 8, 925–933.

Bobinski, M., Wegiel, J., Wisniewski, H.M., Tarnawski, M., Bobinski, M., Reisberg, B., De Leon, M.J., Miller, D.C., 1996. Neurofibrillary pathology—Correlation with hippocampal formation atrophy in Alzheimer disease. Neurobiol. Aging 17 (6), 909–919.

Carmichael, O.T., Aizenstein, H.A., Davis, S.W., Becker, J.T., Thompson, P.M., Meltzer, C.C., Liu Y., 2004. Atlas-based hippocampus segmentation in Alzheimer's disease and mild cognitive impairment. Technical report, Carnegie Mellon University Robotics Institute.

Chard, D.T., Parker, G.J., Griffin, C.M., Thompson, A.J., Miller, D.H., 2002. The reproducibility and sensitivity of brain tissue volume measurements derived from an spm-based segmentation methodology. J. Magn. Reson. Imaging 15 (3), 259–267 (March).

Chen, M., 1999. 3-D Deformable Registration Using a Statistical Atlas with pplications in Medicine, PhD thesis. Robotics Institute, Carnegie Mellon University, Pittsburgh, PA (October).

Chetelat, G., Baron, J.-C., 2003. Early diagnosis of Alzheimer's disease: contribution of structural neuroimaging. NeuroImage 18, 525–541.

Christensen, G.E., Joshi, S.C., Miller, M.I., 1997. Volumetric transformation of brain anatomy. IEEE Trans. Med. Imaging 16 (6), 864–877 (December).

Collins, D., Holmes, C., Peters, T., Evans, A., 1995. Automatic 3d model-based neuroanatomical segmentation. Hum. Brain Mapp. 3 (3), 190–208.

Convit, A., De Leon, M.J., Tarshish, C., De Santi, S., Tsui, W., Rusinek, H., George, A., 1997. Specific hippocampal volume reductions in individuals at risk for Alzheimer's disease. Neurobiol. Aging 18 (2), 131–138 (March-April).

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

Dawant, B.M., Hartmann, S.L., Thirion, J.P., Maes, F., Vandermeulen, D., Demaerel, P., 1999. 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 Trans. Med. Imaging 18 (10), 909–916 (October).

de Leon, M.J., Golomb, J., George, A.E., Convit, A., Tarshish, C.Y., McRae, T., De Santi, S., Smith, G., Ferris, S.H., Noz, M., et al., 1993. The radiologic prediction of Alzheimer disease: the atrophic hippocampal formation. Am. J. Neuroradiol. 14 (4), 897–906 (July–August).

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

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

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

Freeborough, P.A., Fox, N.C., Kitney, R.I., 1997. Interactive algorithms for the segmentation and quantitation of 3-d MRI brain scans. Comput. Methods Prog. Biomed. 53 (1), 15–25 (May).

Frisoni, G.B., 2001. Structural imaging in the clinical diagnosis of Alzheimer's disease: problems and tools. J. Neurol., Neurosurg. Psychiatry 70 (6), 711–718 (September).

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

Geuze, E., Vermetten, E., Bremner, J.D., 2004. MR-based in vivo hippocampal volumetrics: 1. review of methodologies currently employed. Mol. Psychiatry, 1–13 (August 31).

Hogan, R.E., Wang, L., Bertrand, M.E., Willmore, L.J., Bucholz, R.D., Nassif, A.S., Csernansky, J.G., 2004. MRI-based high-dimensional hippocampal mapping in mesial temporal lobe epilepsy. Brain 127 (8), 1731–1740 (August).

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved Methods for the Registration and Motion Correction of Brain Images. Technical report, Oxford Centre for Functional Magnetic Resonance Imaging of the Brain.

Jack Jr., C.R., Theodore, W.H., Cook, M., McCarthy, G., 1995. MRI based hippocampal volumetrics: data acquisition, normal ranges, and optimal protocol. Magn. Reson. Imaging 13, 1057–1064.

Kelemen, A., Szekely, G., Gerig, G., 1999. Elastic model-based segmentation of 3-d neuroradiological data sets. IEEE Trans. Med. Imaging 18 (10), 828–839 (October).

Kikinis, R., Portas, C.M., Donnino, R.M., Jolesz, F.A., Shenton, M.E., Iosifescu, D.V., McCarley, R.W., Saiviroonporn, P., Hokama, H.H., Robatino, A., Metcalf, D., Wible, C.G., 1996. A digital brain atlas for surgical planning, model-driven segmentation, and teaching. IEEE Trans. Vis. Comput. Graph. 2 (3), 232–241 (September).

Klemencic, J., Valencic, V., Pecaric, N., 2001. Deformable contour based algorithm for segmentation of the hippocampus from MRI. Proceedings of the International Conference on Computer Analysis of Images and Patterns, pp. 298–308.

Kordower, J.H., Chu, Y., Stebbins, G.T., DeKosky, S.T., Cochran, E.J., Bennett, D., Mufson, E.J., 2001. Loss and atrophy of layer II entorhinal cortex neurons in elderly people with mild cognitive impairment. Ann. Neurol. 49 (2), 202–213 (February).

Leemput, K.V., Maes, F., Vandermeulen, D., Suetens, P., 1999. Automated model-based tissue classification of MR images of the brain. IEEE Trans. Med. Imaging 18, 897–908.

Lopez, O.L., Becker, J.T., Klunk, W., Saxton, J., Hamilton, R.L., Kaufer, D.I., Sweet, R., Cidis Meltzer, C., Wisniewski, S., Kamboh, M.I., DeKosky, S.T., 2000a. Research evaluation and diagnosis of probable Alzheimer's disease over the last two decades: I. Neurology 55, 1854–1862.

Lopez, O.L., Becker, J.T., Klunk, W., Saxton, J., Hamilton, R.L., Kaufer, D.I., Sweet, R., Cidis Meltzer, C., Wisniewski, S., Kamboh, M.I., DeKosky, S.T., 2000b. Research evaluation and diagnosis of probable Alzheimer's disease over the last two decades: II. Neurology 55, 1863–1869.

Pantel, J., Cretsinger, K., Keefe, H., 1998. Hippocampus tracing guidelines. Available at: http://www.psychiatry.uiowa.edu/ipl/pdf/hippocampus.pdf.

Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Kokmen, E., Tangelos, E.G., 1997. Aging, memory, and mild cognitive impairment. Int. Psychogeriatr. 9 (Suppl. 1), 65–69.

Pinheiro, J.C., Bates, D.M., 2000. Mixed-effects models in S and S-PLUS. Statistics and Computing, Springer.

Pitiot, A., Toga, A.W., Thompson, P.M., 2002. Adaptive elastic segmentation of brain MRI via shape-model-guided evolutionary programming. IEEE Trans. Med. Imaging 21 (8), 910–923 (August).

Pizer, S.M., Fritsch, D.S., Yushkevich, P., Johnson, V., Chaney, E., Gerig, G., 1999. Segmentation, registration, and measurement of shape variation via image object shape. IEEE Trans. Med. Imaging, 851–865 (October).

Rohlfing, T., Brandt, R., Menzel, R., Maurer Jr., C.R., 2004. 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).

Rosenthal, R., Rosnow, R.L., Rubin, D.B., 2000. Contrasts and Effect Sizes in Behavioral Research: A Correlational Approach. Cambridge Univ. Press, Cambridge, UK.

Schuff, N., Amend, D., Ezekiel, F., Steinman, S.K., Tanabe, J., Norman, D., Jagust, W., Kramer, J.H., Mastrianni, J.A., Fein, G., Weiner, M.W., 1997. Changes of hippocampal *n*-acetyl aspartate and volume in Alzheimer's disease. A proton MR spectroscopic imaging and MRI study. Neurology 49, 1513–1521.

Shen, D., Moffat, S., Resnick, S.M., Davatzikos, C., 2002. Measuring size and shape of the hippocampus in MR images using a deformable shape model. NeuroImage 15 (2), 422–434 (February).

Smith, S., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17 (3), 143–155.

Thompson, P.M., Hayashi, K.M., de Zubicaray, G., Janke, A.L., Rose, S.E., Semple, J., Hong, M.S., Herman, D., Gravano, D., Doddrell, D.M., Toga, A.W., 2004. Mapping hippocampal and ventricular change in Alzheimer's disease. NeuroImage, 1754–1766 (June).

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., 2002. Automated anatomical labelling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single subject brain. NeuroImage 15, 273–289.

Warfield, S.K., Robatino, A., Dengler, J., Jolesz, F.A., Kikinis, R., 1998. Nonlinear registration and template driven segmentation. Progressive Publishing Alternatives (chapter 4).

Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging 23 (7), 903–921 (July).

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

Webb, J., Guimond, A., Eldridge, P., Chadwick, D., Meunier, J., Thirion, J.P., Roberts, N., 1999. Automatic detection of hippocampal atrophy on magnetic resonance images. Magn. Reson. Imaging 17 (8), 1149–1161 (October).

Woods, R.P., Grafton, S.T., Holmes, C.J., Cherry, S.R., Mazziotta, J.C., 1998. Automated image registration: I. General methods and intrasubject, intramodality validation. J. Comput. Assist. Tomogr. 22, 139–152.

Yoo, T.S. (Ed.), 2004. lnsight Into Images. Insight Software Consortium.