

Automated Segmentation of Hippocampal Subfields From Ultra-High Resolution In Vivo MRI

Koen Van Leemput,^{1,2*} Akram Bakkour,^{1,3} Thomas Benner,¹ Graham Wiggins,¹
Lawrence L. Wald,^{1,4} Jean Augustinack,¹ Bradford C. Dickerson,^{1,3†}
Polina Golland,^{2‡} and Bruce Fischl^{1,2,4†}

ABSTRACT: Recent developments in MRI data acquisition technology are starting to yield images that show anatomical features of the hippocampal formation at an unprecedented level of detail, providing the basis for hippocampal subfield measurement. However, a fundamental bottleneck in MRI studies of the hippocampus at the subfield level is that they currently depend on manual segmentation, a laborious process that severely limits the amount of data that can be analyzed. In this article, we present a computational method for segmenting the hippocampal subfields in ultra-high resolution MRI data in a fully automated fashion. Using Bayesian inference, we use a statistical model of image formation around the hippocampal area to obtain automated segmentations. We validate the proposed technique by comparing its segmentations to corresponding manual delineations in ultra-high resolution MRI scans of 10 individuals, and show that automated volume measurements of the larger subfields correlate well with manual volume estimates. Unlike manual segmentations, our automated technique is fully reproducible, and fast enough to enable routine analysis of the hippocampal subfields in large imaging studies. © 2009 Wiley-Liss, Inc.

KEY WORDS: hippocampal subfields; segmentation; computational methods; MRI; statistical modeling

INTRODUCTION

The hippocampal formation is critical to episodic memory, and is affected by aging, Alzheimer's disease (AD), schizophrenia, epilepsy, and other conditions. The distinct subregions composing the hippocampus

have been implicated in different memory subsystems (Gabrieli et al., 1997; Zeineh et al., 2003), and shown to be differentially affected in normal aging and AD (West et al., 1994; Fukutani et al., 1995; Small et al., 1999; Small et al., 2000). Therefore, the ability to reliably and efficiently detect these subregions using in vivo neuroimaging is of great potential value for both basic neuroscience and clinical research. Such a procedure would enable studying the function and structure of the hippocampus in the living human brain, and how it is affected in normal aging. It would also be an important step in the quest for sensitive, noninvasive biomarkers for early diagnosis and treatment evaluation in AD.

The image resolution of magnetic resonance imaging (MRI) scans has traditionally been too coarse for identification of individual subregions of the hippocampal formation, forcing investigators to treat the hippocampus as a single entity. Recently, however, substantial developments in MRI data acquisition technology have made it possible to acquire images with a remarkably higher resolution and signal-to-noise (S/N) ratio than was previously attainable. Such scans show many fine-scaled features of the brain at an unprecedented level of detail, and offer new opportunities for explicitly studying individual hippocampal subregions, rather than their agglomerate, directly from in vivo MRI (Gabrieli et al., 1997; Small et al., 1999; Zeineh et al., 2000, 2001, 2003; Miller et al., 2005; Mueller et al., 2007; Kirwan et al., 2007; Burggren et al., 2008; Yushkevich et al., 2009).

While the high field MRI technology needed for visualizing hippocampal subfields is becoming increasingly available, a fundamental bottleneck in current imaging studies is that manual interaction is required to draw subfield boundaries in the images. This severely limits the amount of data that can be analyzed, because delineating subfields manually in ultra-high resolution images is an excruciatingly time consuming process (several days for a full hippocampus in our images). Furthermore, manual delineations suffer from intra- and interobserver variability, which confounds subsequent statistical analyses of the results.

Thus, there is a need for computational methods that can reliably generate hippocampal subfield segmentations in large imaging studies in a fully automated and reproducible fashion. In this article, we

¹Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Massachusetts; ²Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Massachusetts; ³Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Massachusetts; ⁴Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Massachusetts

[†]B.C.D., P.G., and B.F. contributed equally to this work.

Grant sponsor: NIH NCR; Grant number: P41-RR14075, R01 RR16594-01A1, NAC P41-RR13218; Grant sponsor: BIRN Morphometric Project; Grant numbers: BIRN002, U24 RR021382; Grant sponsor: NIBIB; Grant numbers: R01 EB001550, R01EB006758, NIMIC U54-EB005149; Grant sponsor: NINDS; Grant numbers: R01 NS052585-01, R01 NS051826; Grant sponsors: MIND Institute, The Autism & Dyslexia Project funded by the Ellison Medical Foundation.

*Correspondence to: Koen Van Leemput, Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Building 149, Room 2301, 13th Street, Charlestown, MA 02129. E-mail: koen@nmr.mgh.harvard.edu

Accepted for publication 9 March 2009

DOI 10.1002/hipo.20615

Published online 29 April 2009 in Wiley InterScience (www.interscience.wiley.com).

propose such a method, and demonstrate its performance on T1-weighted ultra-high resolution MRI scans.

MATERIALS AND METHODS

We use a Bayesian modeling approach, in which we first build an explicit computational model of how an MRI image around the hippocampal area is generated, and subsequently use this model to obtain fully automated segmentations. We start by describing the computational model and how we optimize its model parameters, then explain how we derive automated subfield segmentations from the model, and finally provide implementation details.

Generative Model

Our computational model incorporates a *prior* distribution that makes predictions about where neuroanatomical labels are a priori expected to occur throughout the image area of a subject's MRI scan. This prior is based on a generalization of probabilistic atlases (Ashburner and Friston, 1997, 2005; Van Leemput et al., 1999a; Fischl et al., 2002; Pohl et al., 2006; Shattuck et al., 2008) we recently developed (Van Leemput, (in press)), and is automatically learned from manual segmentations of the hippocampal formation in MRI images of a number of training subjects. A *likelihood* distribution that predicts how a label image, where each voxel is assigned a unique neuroanatomical label, translates into an MRI image, where each voxel has an intensity, completes the model.

Prior: Mesh-based probabilistic atlas

Let $L = \{l_i, i = 1, \dots, I\}$ be a label image with a total of I voxels, with $l_i \in \{1, \dots, K\}$ denoting the one of K possible neuroanatomical labels assigned to voxel i . Our prior models this image as generated by the following process:

- A tetrahedral mesh covering the image domain of interest is defined by the *reference position* of its N mesh nodes $\mathbf{x}^r = \{\mathbf{x}_n^r, n = 1, \dots, N\}$, and by a set of label probabilities $\alpha = \{\alpha_n, n = 1, \dots, N\}$. Node n is associated with a probability vector $\alpha_n = \{\alpha_n^1, \dots, \alpha_n^K\}$, satisfying $\alpha_n^k \geq 0$ and $\sum_k \alpha_n^k = 1$, that governs how frequently each label occurs around that node.
- The mesh is deformed from its reference position by sampling from a Markov random field model regulating the position of the mesh nodes:

$$p(\mathbf{x}) \propto \exp(-U(\mathbf{x})) = \exp\left(-\sum_{t=1}^T U_t(\mathbf{x})\right) \quad (1)$$

where $U_t(\mathbf{x})$ is a penalty for deforming tetrahedron t from its shape in the reference position \mathbf{x}^r , and $U(\mathbf{x})$ is an overall deformation penalty obtained by summing the contributions of all T

tetrahedra in the mesh. We use the penalty proposed in (Ashburner et al., 2000), which goes to infinity if the Jacobian determinant of any tetrahedron's deformation approaches zero, and therefore insures that the mesh topology is preserved. More details about this deformation penalty are given in the Appendix.

- In the deformed mesh with position \mathbf{x} , the probability of having label k in a pixel i with location \mathbf{x}_i is modeled by

$$p_i(k|\mathbf{x}) = \sum_{n=1}^N \alpha_n^k \Phi_n(\mathbf{x}_i), \quad (2)$$

where $\Phi_n(\cdot)$ denotes an interpolation basis function attached to mesh node n that has a unity value at the position of the mesh node, a zero value at the outward faces of the tetrahedra connected to the node and beyond, and a linear variation across the volume of each tetrahedron. Assuming conditional independence of the labels between voxels given the mesh node locations, we finally have $p(L|\mathbf{x}) = \prod_{i=1}^I p_i(l_i|\mathbf{x})$ for the probability of seeing label image L .

Learning the prior from example segmentations

The probabilistic atlas described above is defined by the connectivity of its tetrahedral mesh, its reference position \mathbf{x}^r , and the probabilities of label occurrences α . Using the techniques developed in (Van Leemput, (in press)), we automatically learn these properties from a set of example segmentations. The learning involves maximizing the probability with which the atlas model would generate the example segmentations, or, equivalently, minimizing the number of bits needed to encode them. As shown in (Van Leemput, (in press)), this procedure automatically yields sparse atlas representations that explicitly avoid overfitting to the training data, and that are therefore better at predicting the neuroanatomy in new subjects than conventional probabilistic atlases.

The segmentations we use for atlas computation are based on manual delineations of the hippocampal subfields in ultra-high resolution T1-weighted MRI scans of a number of different subjects. These delineations include the fimbria, presubiculum, subiculum, CA1, CA2/3, and CA4/DG fields, as well as choroid plexus, hippocampal fissure, and inferior lateral ventricle, as shown in Figure 1. Because the hippocampal formation covers only a small part of the images, we define a cuboid region of interest (ROI) encompassing all the structures of interest in all subjects after affine registration, and model the segmentations within this ROI only. (More details about the manual segmentation protocol and the definition of the ROI are given below.) Prior to atlas computation, voxels inside the ROI not belonging to one of the manually delineated subregions are automatically labeled as white matter, gray matter, or CSF using a brain MRI tissue classification algorithm (Van Leemput et al., 1999b), as these tissues provide useful additional information about the global anatomy in and around the hippocampal formation.

An example of the prior, learned from hippocampal labels in nine subjects, is shown in Figure 2.

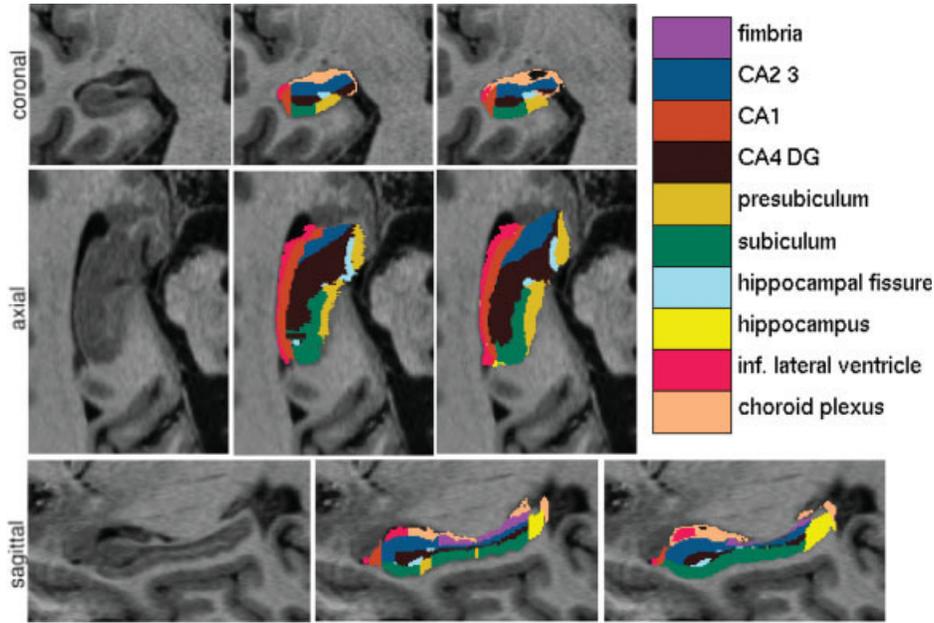


FIGURE 1. From left to right: cross-sectional slices of an ultra-high resolution MRI scan, manual delineation of the hippocampal subfields and corresponding automated segmentation.

Likelihood: Imaging model

For the likelihood distribution, we model an intensity image $Y = \{y_i, i = 1, \dots, I\}$ as generated from a label image L by drawing the intensity in each voxel independently from a parametric distribution associated with its label:

$$p(Y|L, \theta) = \prod_{i=1}^I p_i(y_i|\theta),$$

where θ denotes the likelihood distribution parameters. We model each of the distributions $p_k(y|\theta)$ as a normal (Gaussian) distribution

$$N(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right)$$

with a mean μ and a variance σ^2 . Reflecting the fact that there is little intensity contrast between the cerebral gray matter and the hippocampal subfields subiculum, presubiculum, CA1, CA2/3, and CA4/DG in our images, we consider them part of a global “gray matter” (GM) tissue type with a shared mean and variance. We similarly consider the cerebral white matter and the fimbria as part of a global “white matter” (WM) tissue type, and the hippocampal fissure, the inferior lateral ventricle, and all other CSF as part of a “CSF” tissue type. The only remaining label is that for choroid plexus (CP), CSF producing cells that appear brighter than CSF in T1-weighted images, for which we assume a separate normal distribution.

In summary,

$$p_k(y|\theta) = N\left(y|\mu_{G(k)}, \sigma_{G(k)}^2\right) \quad \forall k$$

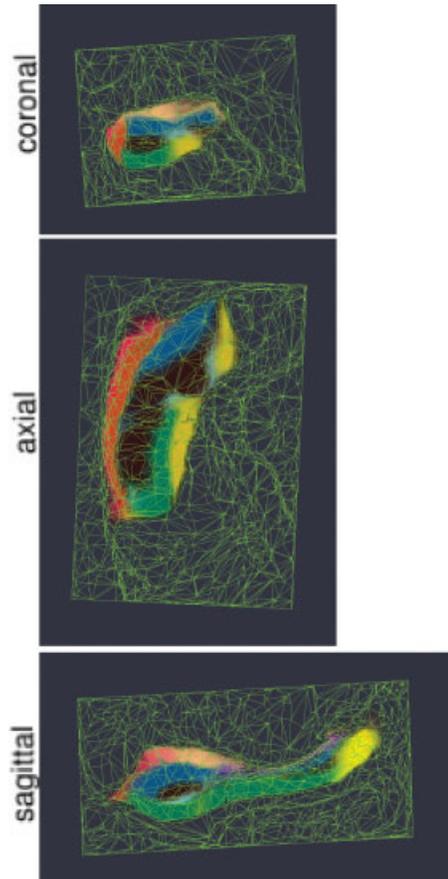


FIGURE 2. Mesh-based probabilistic atlas derived from manual delineations in nine subjects, warped onto the 10th subject which is shown in Figure 1. The structure-wise probabilities have been color-coded for visualization purposes; only 7 structures of interest are shown.

where the notation $G(k)$ is used to indicate which of the four global tissue types {GM, WM, CSF, CP} label k belongs to. The likelihood model parameters are therefore the mean and variance of the normal distribution associated with each global tissue type G : $\boldsymbol{\theta} = \{\{\mu_G, \sigma_G^2\}\}$. To complete the model, we assume a uniform prior on these parameters: $p(\boldsymbol{\theta}) \propto 1$.

Parameter Optimization

With the generative model in place, segmentation of an image Y can proceed by first estimating the model parameters from the data. Assessing the Maximum A Posteriori (MAP) parameter estimates $\{\tilde{\mathbf{x}}, \tilde{\boldsymbol{\theta}}\}$ involves maximizing $p(\mathbf{x}, \boldsymbol{\theta}|Y) \propto p(Y|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x})$, which is equivalent to maximizing

$$\begin{aligned} \log[p(Y|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x})] &= \log \left[\sum_L p(Y|L, \boldsymbol{\theta})p(L|\mathbf{x}) \right] + \log p(\mathbf{x}) \\ &= \sum_{i=1}^I \left(\log \left[\sum_G N(y_i|\mu_G, \sigma_G^2)p_i(G|\mathbf{x}) \right] \right) \\ &\quad + \log p(\mathbf{x}), \end{aligned}$$

where in the last step a prior for each global tissue type G was introduced that is obtained by summing over all its constituent labels: $p_i(G|\mathbf{x}) = \sum_{k \in G} p_i(k|\mathbf{x})$.

For the optimization, we use a *Generalized Expectation-Maximization* (GEM) algorithm (Dempster et al., 1977). Noticing that evaluating the objective function involves taking the logarithm of a sum in each voxel, we exploit Jensen's inequality to construct a lower bound for it. Given some estimate $\{\tilde{\mathbf{x}}, \tilde{\boldsymbol{\theta}}\}$ for the parameters, and defining

$$\tilde{W}_i^G \equiv \frac{N(y_i|\tilde{\mu}_G, \tilde{\sigma}_G^2)p_i(G|\tilde{\mathbf{x}})}{\sum_{G'} N(y_i|\tilde{\mu}_{G'}, \tilde{\sigma}_{G'}^2)p_i(G'|\tilde{\mathbf{x}})} \quad (3)$$

as a statistical classification that associates each voxel with each of the global tissue types based on that estimate, we form a lower bound $Q(\mathbf{x}, \boldsymbol{\theta}|\tilde{\mathbf{x}}, \tilde{\boldsymbol{\theta}})$:

$$\begin{aligned} Q(\mathbf{x}, \boldsymbol{\theta}|\tilde{\mathbf{x}}, \tilde{\boldsymbol{\theta}}) &= \sum_{i=1}^I \left(\sum_G \tilde{W}_i^G \log \left[\frac{N(y_i|\mu_G, \sigma_G^2)p_i(G|\mathbf{x})}{\tilde{W}_i^G} \right] \right) \\ &\quad + \log p(\mathbf{x}) \\ &\leq \sum_{i=1}^I \left(\log \left[\sum_G \left(\frac{N(y_i|\mu_G, \sigma_G^2)p_i(G|\mathbf{x})}{\tilde{W}_i^G} \right) \tilde{W}_i^G \right] \right) \\ &\quad + \log p(\mathbf{x}) = \log[p(Y|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x})]. \end{aligned}$$

This lower bound touches the objective function at the parameter value $\{\mathbf{x}, \boldsymbol{\theta}\} = \{\tilde{\mathbf{x}}, \tilde{\boldsymbol{\theta}}\}$, i.e., $Q(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\theta}}|\tilde{\mathbf{x}}, \tilde{\boldsymbol{\theta}}) = \log[p(Y|\tilde{\mathbf{x}}, \tilde{\boldsymbol{\theta}})p(\tilde{\mathbf{x}})]$, as can be checked by substituting Eq.

(3) into the definition of the lower bound and noting that $\sum_G \tilde{W}_i^G = 1$.

Our GEM algorithm optimizes the objective function by iteratively constructing the lower bound at the current parameter estimate $\{\tilde{\mathbf{x}}, \tilde{\boldsymbol{\theta}}\}$, updating the estimate to improve the lower bound, and repeating this process until convergence. Since the lower bound always touches the objective function at the current parameter estimate, this procedure guarantees that the objective function is improved with each new iteration, unless a local maximum or saddle point has already been reached. In our implementation, we keep the atlas warp \mathbf{x} fixed for several iterations and repeatedly optimize the lower bound with respect to the normal distribution parameters $\boldsymbol{\theta}$ only, until no further improvements in $\boldsymbol{\theta}$ can be made. Subsequently, we keep the normal distribution parameters fixed for one iteration and optimize the lower bound with respect to the atlas warp only, after which the whole process is repeated, until convergence. Details for the optimization of each parameter set are given below.

Optimizing the lower bound with respect to the normal distribution parameters

The normal distribution parameters that optimize the lower bound for a given atlas warp are given by the following closed form expressions:

$$\mu_G \leftarrow \frac{\sum_{i=1}^I \tilde{W}_i^G y_i}{\sum_{i=1}^I \tilde{W}_i^G}, \quad \sigma_G^2 \leftarrow \frac{\sum_{i=1}^I \tilde{W}_i^G (y_i - \mu_G)^2}{\sum_{i=1}^I \tilde{W}_i^G}, \quad \forall G. \quad (4)$$

The updated normal distribution parameter estimates are thus the sample mean and variance of the voxels classified to each global tissue type, where the classification is based upon the current atlas warp and normal distribution parameters [Eq. (3)].

Optimizing the lower bound with respect to the atlas warp

Optimizing the lower bound with respect to \mathbf{x} reduces to optimizing

$$\sum_{i=1}^I \sum_G \tilde{W}_i^G \log p_i(G|\mathbf{x}) + \log p(\mathbf{x}) \quad (5)$$

and thus involves deforming the atlas towards the current classification, similar to the schemes proposed in (Pohl et al., 2006; D'Agostino et al., 2006). For the optimization, we use an iterative strategy similar to the one used by the Levenberg-Marquardt algorithm for solving nonlinear least-squares problems (Press et al., 1992). In each iteration, we determine an incremental update $\delta\mathbf{x}$ by solving the sparse linear system

$$[H + \lambda \text{diag}(H)] \cdot \delta \mathbf{x} = g, \quad (6)$$

where g and H denote the gradient and the Hessian of Eq. (5) at the current estimate \mathbf{x} , respectively. Both g and H are given in analytical form through the interpolation model of Eq. (2) and the deformation model of Eq. (1). For convenience we ignore second derivate terms in the computation of H . The parameter λ is a regularization parameter that allows the method to automatically vary between an efficient inverse-Hessian and a slower but more robust gradient descent approach. In each iteration, Eq. (6) is solved for $\delta \mathbf{x}$, and Eq. (5) is evaluated at the position $\mathbf{x} + \delta \mathbf{x}$. If Eq. (5) is improved, $\mathbf{x} + \delta \mathbf{x}$ is taken as the new estimate of \mathbf{x} , λ is decreased, and a new iteration is started; otherwise the current value of \mathbf{x} is retained, λ is increased, and Eq. (6) is solved for a new trial step. The iterations are stopped when the improvement in Eq. (5) between two consecutive iterations becomes negligible.

Hippocampal Subfield Segmentation

Once we have an estimate of the model parameters $\{\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\}$, we use it to obtain an approximation to the MAP anatomical labeling. Assuming $p(\mathbf{x}, \boldsymbol{\theta} | Y)$ is strongly peaked at $\{\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\}$, the optimal segmentation is given by

$$\begin{aligned} \hat{L} &= \arg \max_L p(L|Y) = \arg \max_L \int \int p(L|Y, \mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}, \boldsymbol{\theta} | Y) d\mathbf{x} d\boldsymbol{\theta} \\ &\approx \arg \max_L p(L|Y, \hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}) \\ &= \arg \max_{\{l_i, i=1, \dots, I\}} \prod_{i=1}^I p_i(l_i | y_i, \hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}) \end{aligned}$$

which is obtained by assigning each voxel to the label with the highest posterior probability

$$p_i(k | y, \hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}) \propto N(y | \hat{\mu}_{G(k)}, \hat{\sigma}_{G(k)}^2) p_i(k | \hat{\mathbf{x}}).$$

Implementation

As mentioned previously, we only consider the area within a ROI around the hippocampal formation in each image. To this end, we defined a cuboid ROI of size $94 \times 66 \times 144$ voxels encompassing the hippocampus in the ICBM 452 template, which is an average of T1-weighted MRIs of normal young adult brain. We automatically align this ROI to each image under study using an affine Mutual Information based registration technique (Wells et al., 1996; Maes et al., 1997), by first aligning the whole template image covering the entire brain, followed by a registration of the ROI only. In order to allow for optimal alignment of the hippocampus in the latter registration phase, we manually defined a rough outline around the hippocampus in the ROI, and replaced the intensity of voxels outside it with zero values. After ROI alignment, we compute

(in the atlas construction phase) and apply (in the segmentation phase) atlas meshes only in the area covered by the cuboid ROI in each image.

Before segmenting the ROI, we automatically correct the data for MRI intensity inhomogeneities using a previously developed technique (Van Leemput et al., 1999b). We initialize the model parameter estimation algorithm by setting the initial position of the mesh nodes \mathbf{x} to the mesh's reference position \mathbf{x}^r and by using Eq. (4) to estimate initial values for the likelihood distribution parameters $\boldsymbol{\theta}$. In order to reduce the risk of getting trapped in a local optimum, we employ a multi-resolution optimization strategy, in which the atlas mesh is subject to a gradually decreasing amount of spatial smoothing. Specifically, we use a three-level strategy, where in the first two levels the label probabilities of each mesh node are averaged with those of the nodes nearby using a Gaussian weighting kernel with a standard deviation (SD) of 2 and 1 times the voxel size, respectively, and the last multi-resolution level uses the original atlas mesh without smoothing.

EXPERIMENTS

We performed experiments on ultra-high resolution MRI data collected as part of an ongoing imaging study assessing the effects of normal aging and AD on brain structure. Using a prototype custom-built 32-channel head coil with a 3.0T Siemens Trio MRI system (Wiggins et al., 2006), we acquired images via an optimized high-resolution MPRAGE sequence that enables $380 \mu\text{m}$ in-plane resolution (TR/TI/TE = 2530/1100/5.39 ms, FOV = 448, FA = 7° , 208 slices acquired coronally, thickness = 0.8 mm, acquisition time = 7.34 min). To increase the signal-to-noise ratio, five acquisitions were collected and motion-corrected to obtain a single resampled (to $380 \mu\text{m}$ isotropic) high contrast volume that covers the entire medial temporal lobe.

Using a protocol developed specifically for this purpose, the subfields of the right hippocampus were manually delineated in images of 10 subjects (six younger and four older cognitively normal individuals; five male and five female; age range 22–89). The delineation protocol is currently being validated and refined and is used here in its initial form for the primary purpose of the development and reliability assessment of the automated segmentation technique that is the main focus of the present study. The hippocampal formation is viewed in the coronal plane starting at the first slice of the body (caudal to the uncus portion). The border between the presubiculum and the subiculum is drawn by extending a vertical line down from the medial most edge of the fimbria. The medial border of presubiculum is a vertical line at the dorsomedial crown of the parahippocampal gyrus. The medial border of CA1 is drawn by connecting the points of highest curvature on the dorsolateral and ventrolateral edges of the hippocampus. This also constitutes the lateral boundary for CA2/3, CA4/DG and subiculum. The dorsal boundary of the subiculum is drawn along

the subicular clouds and/or the edge of the hippocampal fissure, if visible. This is also the ventral border of CA4/DG. The lateral boundary of CA1 and the dorsal boundary of CA2/3 is the border between hippocampus and the temporal horn of the lateral ventricle. Along the entire extent of the hippocampal formation, the dorsolateral white matter is labeled fimbria/alveus. A line through the intersection of two tangents at the points of maximum curvature mentioned above and perpendicular to the medial boundary of CA1 constitutes the ventral border of CA2/3 and dorsal boundary for CA4/DG. This protocol is applied to every slice moving caudally until the fornix is fully formed, at which point the entire hippocampal formation is labeled as simply “hippocampus.” In the hippocampal head, the same general procedure is followed. Using this procedure, each hippocampus takes approximately 2.5 days to label.

We evaluated our automated segmentation algorithm using a leave-one-out cross-validation strategy: we built an atlas mesh from the manual delineations in nine of the 10 subjects, and used this to segment the image of the remaining subject. We repeated this process for each of the 10 subjects, and compared the automated segmentation results with the corresponding manual delineations. Towards the tail of the hippocampus, the manual delineations no longer discern between the different subfields, but rather lump everything together as simply “hippocampus,” as shown in Figure 1. Since the volume of this “catch-all” label varies widely between subjects (from 5 to 17% of the total hippocampal volume), and information on its starting point is not available to the automated algorithm, voxels that were labeled as such in either the automated or manual segmentation were not included in the comparisons.

For each of seven structures of interest (fimbria, CA1, CA2/3, CA4/DG, presubiculum, subiculum, and hippocampal fissure), we calculated the Dice overlap coefficient, a widely used segmentation evaluation metric that is defined as the volume of overlap between the automated and manual segmentation divided by their mean volume. Since the Dice overlap coefficient tends to attain higher values for larger structures than for smaller ones, we also calculated the average distance between the boundary of each structure’s manual segmentation and the boundary of the corresponding automated segmentation. Furthermore, since we are ultimately interested in using the automated method for detecting subtle morphological changes in hippocampal subfields, we also evaluated how well differences in subfield volumes between subjects, as detected by the manual delineations, were reflected in the automated segmentations. To this end, we performed a linear regression on the absolute volumes detected by both methods, calculating Pearson’s correlation coefficient for each structure.

To place the automated method’s segmentation performance in the context of human intrarater variability, two consecutive coronal slices in the midbody of the hippocampus in a randomly chosen subset of five subjects (three younger and two older subjects; two male and three female; age range 22–89) were re-delineated twice by the same human operator who performed the original segmentations. We then calculated the Dice overlap coefficients between each structure’s automated segmen-

tation in these two slices and the corresponding manual segmentations, and compared that to the Dice overlap between the manual delineations themselves.

RESULTS

The computational segmentation process for each of the 10 subjects was fully automated and took about 5 h per subject on a 2.83 GHz Intel Xeon E5440 processor. A qualitative comparison between the automated segmentation result in one subject and the corresponding manual delineation is shown in Figure 1.

The top left of Figure 3 shows, for each of the structures of interest, the average Dice overlap measure between the automated and manual segmentation in all 10 subjects, along with error bars that indicate the standard errors around the mean. As expected, larger structures score better than do smaller ones: CA2/3 and subiculum, the largest structures with an average size of 935 and 537 mm³, respectively, have an average Dice coefficient of around 0.74, whereas CA4/DG (on average 526 mm³) and presubiculum (on average 431 mm³) score around 0.68 each, and the smaller CA1 (on average 340 mm³) has a Dice coefficient of around 0.62. Automated segmentation of the fimbria and the hippocampal fissure, the smallest structures with an average volume of only 81 mm³ each, is considerably more challenging, with a Dice coefficient of around 0.51 and 0.53, respectively.

The bottom left of Figure 3 shows the mean and standard errors of the average distance between the boundary of each structure’s manual segmentation and the boundary of the corresponding automated segmentation. As can be seen, this evaluation metric is less dependent on the structures’ size than the Dice overlap coefficient. The average boundary error is below the voxel size of 380 μm for most structures, including the fimbria. Exceptions are CA1, which has an average boundary error of about 1.17 times the voxel size, and the hippocampal fissure, with an average error almost twice the voxel size.

The top right of Figure 3 shows, for each structure, the volume differences between the automated and manual segmentations relative to their mean volumes. Regarding Pearson’s correlation coefficient, the automatically calculated volumes of CA2/3 and CA4/DG are strongly correlated with the manual ones, with a correlation coefficient of 0.91 ($P \leq 0.0002$) and 0.83 ($P \leq 0.0028$), respectively. Subiculum also correlates to some degree (correlation coefficient 0.60), although the correlation does not reach statistical significance ($P \leq 0.066$). Interestingly, despite the hippocampal fissure’s relatively low Dice overlap measure and large average boundary error, its automated volume measurements correlate better with the manual ones than do some structures with much better segmentation evaluation metrics (correlation coefficient 0.50, $P \leq 0.14$). The relatively poor segmentation evaluation scores are apparently caused by a systematic underestimation of the volume of the hippocampal fissure by the automated method.

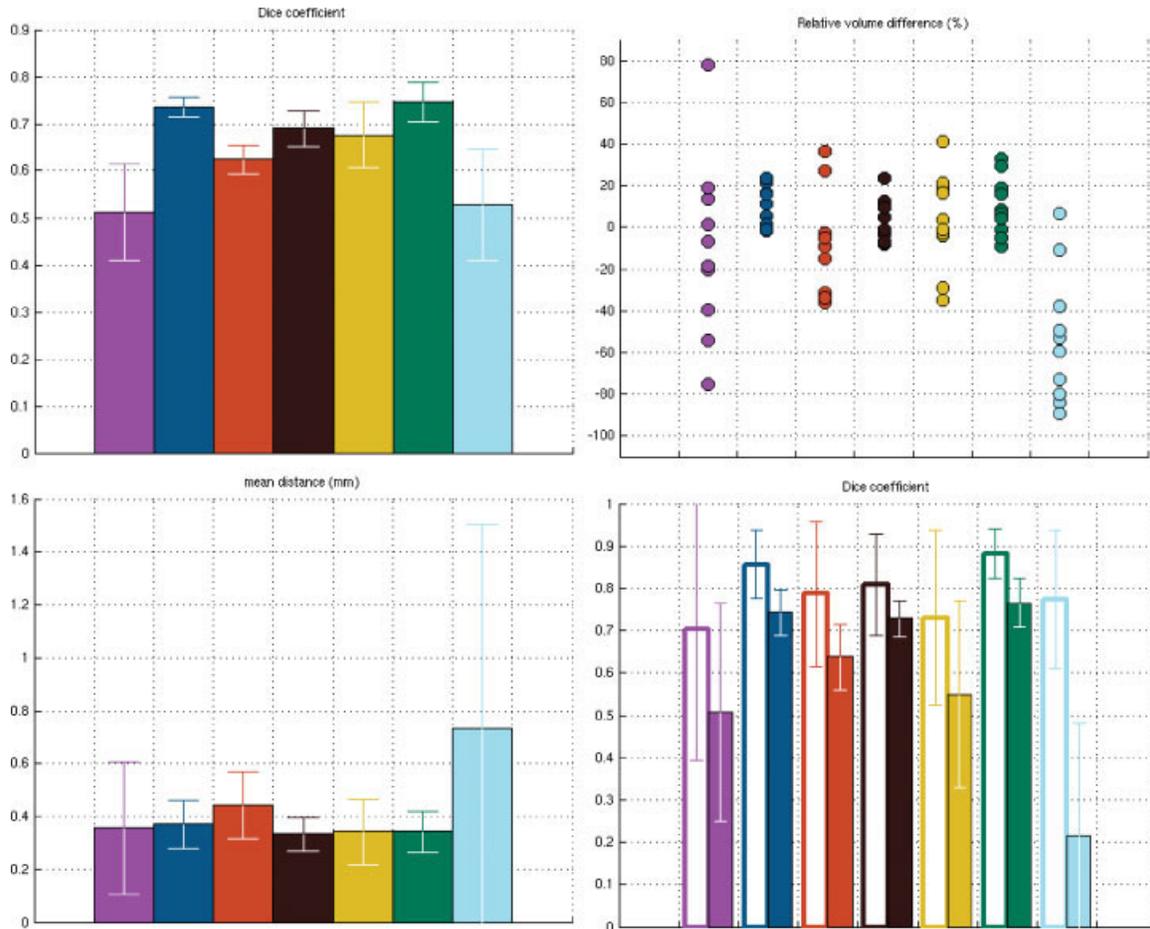


FIGURE 3. Dice overlap measures (top left), average boundary distances (bottom left), and relative volume differences (top right) between automated and manual segmentations in 10 subjects. Also shown (bottom right) are the human intrarater Dice overlap meas-

ures (empty bars) along with automated vs. manual Dice overlaps (filled bars), calculated on two slices in the midbody of the hippocampus in five subjects. The colors are as in Figure 1. Error bars represent standard errors on the mean.

The bottom right of Figure 3 shows, for each structure, the average Dice overlap measure between the automated segmentation and the repeated manual delineations of two slices in the midbody of the hippocampus in five subjects (filled bars), along with the Dice overlap between the repeated manual segmentations themselves (empty bars). Qualitatively, the automated method performs similarly on these selected slices as on the whole volumes, except for the presubiculum (Dice coefficient decreased to 0.55), and the hippocampal fissure (Dice coefficient only 0.2). With respect to the intrarater variability of the human operator, the automated results are systematically more different from the manual segmentations than the manual segmentations from one another, although considerable disagreement is apparent between the latter as well (mean intrarater Dice overlap over all structures: 0.79). With the exception of the hippocampal fissure, the structures that are most difficult to segment reliably by the human operator are also the most difficult for the automated method: the order in which the Dice scores decrease across structures is the same for both methods.

DISCUSSION

In this article, we presented a computational method for automatically segmenting the hippocampal subfields from ultra-high resolution MRI scans, and demonstrated its performance in T1-weighted images of 10 subjects. We showed that automated volume measurements of the larger subfields CA2/3, CA4/DG, and, to a lesser degree, subiculum, correlated well with manual volume estimates. When the degree of spatial correspondence between segmentations was taken into account, considerable disagreement was revealed, both between automated and manual segmentations, as well as between repeated manual measurements themselves. Unlike manual segmentations, our automated technique is fully reproducible, and fast enough to enable routine analysis of the hippocampal subfields in large imaging studies.

Although manual segmentations were considered the gold standard for evaluating our automated algorithm, and repeated

manual segmentations corresponded better with each other than with the automated results, it should be noted that manual measurements suffer from their own type of errors as well. For instance, manual delineations typically appear more consistent in the plane they were drawn in than in other, cross-sectional directions, as can be seen in Figure 1. Moreover, one can not exclude the possibility that a human operator, presented repeatedly with exactly the same image, repeatedly makes the same mistakes. A more comprehensive validation would, therefore, include manual delineations by more than one human operator, performed on several repeat scans for each subject, but such a detailed validation is outside the scope of the current article.

While the results presented here are based on T1-weighted images, the fact that our algorithm automatically learns and adapts to each scan's contrast properties should make it applicable to images acquired with different pulse sequences as well, for example, the T2-weighted protocols used by some other groups for studying the fine details of the hippocampal formation (Mueller et al., 2007; Burggren et al., 2008). Furthermore, generalizing the model and the optimization procedures described here from single-valued to vector-valued intensities is straightforward, allowing the algorithm to directly analyze multi-spectral MRI scans as well. With the continued advancements in MRI acquisition technology, we expect the use of such multi-spectral images, along with further improvements in image resolution and contrast-to-noise ratio, to further increase the segmentation accuracy of our approach in the future.

REFERENCES

- Ashburner J, Friston KJ. 1997. Multimodal image coregistration and partitioning—A unified framework. *Neuroimage* 6:209–217.
- Ashburner J, Friston KJ. 2005. Unified segmentation. *Neuroimage* 26:839–885.
- Ashburner J, Andersson JL, Friston KJ. 2000. Image registration using a symmetric prior—in three dimensions. *Hum Brain Mapp* 9:212–225.
- Burggren AC, Zeineh MM, Ekstrom AD, Braskie MN, Thompson PM, Small GW, Bookheimer SY. 2008. Reduced cortical thickness in hippocampal subregions among cognitively normal apolipoprotein E $\epsilon 4$ carriers. *Neuroimage* 41:1177–1183.
- D'Agostino E, Maes F, Vandermeulen D, Suetens P. 2006. A Unified Framework for Atlas Based Brain Image Segmentation and Registration. *Lecture Notes in Computer Science*. Berlin: Springer-Verlag, pp 136–143.
- Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc* 39:1–38.
- Fischl B, Salat DH, 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 AM. 2002. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron* 33:341–355.
- Fukutani Y, Kobayashi K, Nakamura I, Watanabe K, Isaki K, Cairns NJ. 1995. Neurons, intracellular and extracellular neurofibrillary tangles in subdivisions of the hippocampal cortex in normal ageing and Alzheimer's disease. *Neurosci Lett* 200:57–60.
- Gabrieli JDE, Brewer JB, Desmond JE, Glover GH. 1997. Separate neural bases of two fundamental memory processes in the human medial temporal lobe. *Science* 276:264–266.
- Kirwan CB, Jones CK, Miller MI, Stark CEL. 2007. High-resolution fMRI investigation of the medial temporal lobe. *Hum Brain Mapp* 28:959–966.
- Maes F, Collignon A, Vandermeulen D, Marchal G, Suetens P. 1997. Multi-modality image registration by maximization of mutual information. *IEEE Trans Med Imag* 16:187–198.
- Miller MI, Beg MF, Ceritoglu C, Stark C. 2005. Increasing the power of functional maps of the medial temporal lobe by using large deformation diffeomorphic metric mapping. *Proc Nat Acad Sci* 102:9685–9690.
- Mueller SG, Stables L, Du AT, Schuff N, Truran D, Cashdollar N, Weiner MW. 2007. Measurement of hippocampal subfields and age-related changes with high resolution MRI at 4T. *Neurobiol Aging* 28:719–726.
- Pohl KM, Fisher J, Grimson EL, Kikinis R, Wells WM. 2006. A Bayesian model for joint segmentation and registration. *Neuroimage* 31:228–239.
- Press WH, Flannery BP, Teukolsky SA, Vetterling WT. 1992. *Numerical Recipes in C. The Art of Scientific Computing*, 2nd ed. Cambridge University Press.
- Shattuck DW, Mirza M, Adisetiyo V, Hojatkashani C, Salamon G, Narr KL, Poldrack RA, Bilder RM, Toga AW. 2008. Construction of a 3D probabilistic atlas of human cortical structures. *Neuroimage* 39:1064–1080.
- Small SA, Nava AS, Perera GM, Delapaz R, Stern Y. 2000. Evaluating the function of hippocampal subregions with high-resolution MRI in Alzheimer's disease and aging. *Microsc Res Tech* 51:101–108.
- Small SA, Perera GM, DeLapaz R, Mayeux R, Stern Y. 1999. Differential regional dysfunction of the hippocampal formation among elderly with memory decline and Alzheimer's disease. *Ann Neurol* 45:466–472.
- Van Leemput K. Encoding probabilistic brain atlases using bayesian inference. *IEEE Trans Med Imag* (in press).
- Van Leemput K, Maes F, Vandermeulen D, Suetens S. 1999a. Automated model-based bias field correction of MR images of the brain. *IEEE Transac Med Imag* 18:885–896.
- Van Leemput K, Maes F, Vandermeulen D, Suetens S. 1999b. Automated model-based tissue classification MR images of the brain. *IEEE Transac Med Imag* 18:897–908.
- Wells WM, Viola P, Atsumi H, Nakajima S, Kikinis R. 1996. Multimodal volume registration by maximization of mutual information. *Med Image Anal* 1:35–52.
- West MJ, Coleman PD, Flood DG, Troncoso JC. 1994. Differences in the pattern of hippocampal neuronal loss in normal ageing and Alzheimer's disease. *Lancet* 344:769–772.
- Wiggins GC, Triantafyllou C, Potthast A, Reykowski A, Nittka M, Wald LL. 2006. 32-channel 3 Tesla receive-only phased-array head coil with soccer-ball element geometry. *Magn Reson Med* 56:216–223.
- Yushkevich PA, Avants BB, Pluta J, Das S, Minkoff D, Mechanic-Hamilton D, Glynn S, Pickup S, Liu W, Gee JC, Grossman M, Detre JA. 2009. A high-resolution computational atlas of the human hippocampus from postmortem magnetic resonance imaging at 9.4 T. *Neuroimage* 44:385–398.
- Zeineh MM, Engel SA, Bookheimer SY. 2000. Application of cortical unfolding techniques to functional MRI of the human hippocampal region. *Neuroimage* 11:668–683.
- Zeineh MM, Engel SA, Thompson PM, Bookheimer SY. 2001. Unfolding the human hippocampus with high resolution structural and functional MRI. *Anat Rec (New Anat)* 265:111–120.
- Zeineh MM, Engel SA, Thompson PM, Bookheimer SY. 2003. Dynamics of the hippocampus during encoding and retrieval of face-name Pairs. *Science* 299:577–580.

APPENDIX: DEFORMATION PENALTY

Here, we provide details on the deformation penalty proposed by (Ashburner et al., 2000). Let $(u_{t,j}^r, v_{t,j}^r, w_{t,j}^r)$ and $(u_{t,j}, v_{t,j}, w_{t,j})$ denote the position of the j th vertex of tetrahedron t in the mesh at reference position and after deformation, respectively. The affine mapping of the tetrahedron is then obtain by

$$M = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} u_{t,1} & u_{t,2} & u_{t,3} & u_{t,4} \\ v_{t,1} & v_{t,2} & v_{t,3} & v_{t,4} \\ w_{t,1} & w_{t,2} & w_{t,3} & w_{t,4} \\ 1 & 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} u_{t,1}^r & u_{t,2}^r & u_{t,3}^r & u_{t,4}^r \\ v_{t,1}^r & v_{t,2}^r & v_{t,3}^r & v_{t,4}^r \\ w_{t,1}^r & w_{t,2}^r & w_{t,3}^r & w_{t,4}^r \\ 1 & 1 & 1 & 1 \end{bmatrix}^{-1}$$

The Jacobian matrix \mathbf{J} of this mapping, given by $\mathbf{J} = \begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{bmatrix}$, can be decomposed into two rotations \mathbf{U} and \mathbf{V} and a diagonal matrix \mathbf{S} , using the Singular Value Decomposition (SVD): $\mathbf{J} = \mathbf{USV}^T$, where $\mathbf{S} = \text{diag}(s_1, s_2, s_3)$. Ashburner’s penalty for each tetrahedron is based on its volume and on the singular values s_1 , s_2 , and s_3 , which represent relative stretching in orthogonal directions:

$$U_t(x) = K \cdot V_t^r \cdot \left(1 + \prod_{p=1}^3 s_p \right) \cdot \sum_{p=1}^3 \left(s_p^2 + 1/s_p^2 - 2 \right),$$

where V_t^r denotes the volume of the tetrahedron in the reference position, and K is a regularization parameter (set to 0.01 in our experiments). This penalty can be conveniently calculated without explicitly performing a SVD, as explained in (Ashburner et al., 2000).