Published in final edited form as:
IEEE Trans Med Imaging. 2010 October ; 29(10): 1714–1729. doi:10.1109/TMI.2010.2050897.

A Generative Model for Image Segmentation Based on Label Fusion

Mert R. Sabuncu,
Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA and also with the Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Charlestown, MA 02129 USA

B. T. Thomas Yeo,
Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA

Koen Van Leemput,
Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA, and with the Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Charlestown, MA 02129 USA, and also with the Department of Information and Computer Science, Aalto University School of Science and Technology, FI-00076 Aalto, Finland

Bruce Fischl, and
Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA and also with the Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Charlestown, MA 02129 USA

Polina Golland
Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA

Abstract

We propose a nonparametric, probabilistic model for the automatic segmentation of medical images, given a training set of images and corresponding label maps. The resulting inference algorithms rely on pairwise registrations between the test image and individual training images. The training labels are then transferred to the test image and fused to compute the final segmentation of the test subject. Such label fusion methods have been shown to yield accurate segmentation, since the use of multiple registrations captures greater inter-subject anatomical variability and improves robustness against occasional registration failures. To the best of our knowledge, this manuscript presents the first comprehensive probabilistic framework that rigorously motivates label fusion as a segmentation approach. The proposed framework allows us to compare different label fusion algorithms theoretically and practically. In particular, recent label fusion or multiatlas segmentation algorithms are interpreted as special cases of our framework. We conduct two sets of experiments to validate the proposed methods. In the first set of experiments, we use 39 brain MRI scans—with manually segmented white matter, cerebral

Correspondence to: Mert R. Sabuncu.

M. R. Sabuncu and B. T. T. Yeo contributed equally to this work.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

cortex, ventricles and subcortical structures—to compare different label fusion algorithms and the widely-used FreeSurfer whole-brain segmentation tool. Our results indicate that the proposed framework yields more accurate segmentation than FreeSurfer and previous label fusion algorithms. In a second experiment, we use brain MRI scans of 282 subjects to demonstrate that the proposed segmentation tool is sufficiently sensitive to robustly detect hippocampal volume changes in a study of aging and Alzheimer's Disease.

Index Terms

Image parcellation; image registration; image segmentation

I. Introduction

This paper investigates a probabilistic modeling framework to develop automatic segmentation tools that delineate anatomical regions of interest in a novel medical image scan. The objective is to learn a segmentation protocol from a collection of training images that have been manually labeled by an expert. This protocol is then employed by the algorithm to automatically segment a new (test) image. Such supervised segmentation tools are commonly used in many medical imaging applications, including surgical planning [27] and the study of disease progression, aging or healthy development [23], [50], [74]. As an application domain, this paper focuses on magnetic resonance (MR) imaging of the brain. However, most of the ideas we discuss here can be easily extended to other modalities and applications, particularly with the recent development of fast algorithms for pairwise registration in other imaging domains [71], [73]. We will thus consider the problem of segmenting the MRI volume scan of a novel subject, based on other subjects' MRI scans that have been delineated by an expert.

Early MR segmentation algorithms mainly dealt with the problem of tissue classification, where local image intensity profiles contain a significant amount of the relevant information [10], [15], [41], [68]. A detailed parcellation of the brain anatomy into structurally or functionally defined regions of interest (ROI) typically requires supervision, commonly in the form of labeled training data, since the local appearance of most such structures is not easily distinguishable [23], [30], [40], [51], [72]. The training data is commonly obtained via a time-consuming and/or expensive procedure such as manual delineation, histology or functional localization experiments [22], [48], [60]. Automating the painstaking procedure of labeling improves the reliability and repeatability of the study, while allowing for the analysis of large pools of subjects.

One of the simplest ways to automatically segment an image using a single training dataset is to perform a nonrigid, dense registration between the labeled image and test image. The resulting warp can then be used to map the training labels onto the coordinates of the test image [16], [26], [31], [43]. The quality of such a registration-based approach is limited by the accuracy of the pairwise registration procedure and the anatomical similarity between the labeled and test subjects.

To reduce the bias due to the labeled subject and to model anatomical variability, multiple subjects can be employed in the training phase. A common method is to use a parametric model to summarize the training data in a common coordinate system [8], [24], [29], [36], [48], [62]–[64], [72]. In this approach the training data are co-registered to compute probability maps that encode the prior probability of observing a particular label at each point in the common (atlas) coordinates. The test subject is then normalized to the atlas coordinates through a pairwise registration with a template image that represents the average

subject. This registration can be completed as a preprocessing step, or can be unified with the segmentation procedure, as in [8], [18], [48], [72]. Once the test subject is spatially normalized, one can use a variety of models of shape and appearance to devise a segmentation algorithm. Traditionally, generative models have been popular, where simple conditionally independent Gaussian models are used for appearance [24], [48]. More sophisticated shape models that encourage certain topological properties have also been proposed to improve segmentation quality [47], [62], [70].

Rather than relying on a single atlas coordinate system, an alternative strategy is to register each training subject to the test subject *separately*. Similar to the registration-based segmentation approach, these pairwise warps are then used to transfer the training labels into the coordinate system of the test subject. Given these transferred labels, segmentation has been commonly formulated as a label fusion problem [30], [42], [51]–[53], [58]. Intriguingly, Majority Voting, which is probably the simplest label fusion method, has been demonstrated to yield powerful segmentation tools [1], [30], [51], [52]. Empirical results in these studies suggest that errors in the manual labeling and registration procedures are reduced during label fusion, resulting in accurate segmentation. Recent work has shown that weighted averaging can be used to improve segmentation quality [6], [32], [53], [58]. The underlying intuition is that training subjects more similar to the test subject should carry more weight during label fusion. The practical advantages of various strategies based on this idea have recently been demonstrated [6]. These strategies include using the whole image to determine a single, global weight for each training subject [6], employing local image intensities for locally adapting the weights [6], [32] and using estimates of classifier performance in determining a weight for each label and training subject [53], [67]. To the best of our knowledge, however, none of the label fusion segmentation methods are derived within a probabilistic framework that explicitly models the relationship between the training data and test subject.

Label fusion methods offer two main advantages: 1) across-subject anatomical variability is better captured than in a single atlas, which can be viewed as a parametric model that typically uses single mode distributions (e.g., Gaussian) to encode anatomical appearance, and 2) multiple registrations improve robustness against occasional registration failures. The main drawback is the computational burden introduced by the multiple registrations and information fusion from the entire training data. To alleviate this problem, various heuristics for selecting only a subset of the training subjects for use in registration and label fusion have been proposed [1]. In the atlas construction literature, there has also been recent work on estimating multiple templates from a set of images [2], [11], [54], [66]. We believe that this approach can also be used to reduce the computational cost of label fusion by summarizing the training data. We leave this issue for future work.

The central contribution of this paper is to propose and investigate a generative model that leads to label fusion style image segmentation algorithms. Within the proposed framework, we derive several methods that combine transferred training labels into a single segmentation estimate. Using a dataset of 39 brain MRI scans and corresponding label maps obtained from an expert, we experimentally compare these segmentation algorithms. Additionally, we compare against other benchmarks including FreeSurfer's whole brain segmentation tool, which has been widely used in a large number of studies [69], and STAPLE [67], a method that combines multiple segmentation estimates based on a probabilistic performance model. Our results suggest that the proposed framework yields accurate and robust segmentation tools that can be employed on large multisubject datasets. In a second experiment, we used one of the proposed segmentation algorithms to compute hippocampal volumes in MRI scans of 282 subjects. A comparison of these measurements across clinical groups indicate that the proposed algorithm is sufficiently sensitive to

robustly detect hippocampal volume differences associated with aging and early Alzheimer's Disease.

The generative model described in this paper is an extension of the preliminary ideas we presented in recent conference papers [55], [57]. The present paper offers detailed derivations, discussions and experiments that were not contained in those papers. The remainder of the paper is organized as follows. Sections II and III present the generative model and its instantiation, respectively. In Section IV, we develop several label fusion style segmentation algorithms based on the proposed generative model. Section V presents empirical results. In Section VI, we discuss the contributions of the paper along with the drawbacks of the proposed algorithms, while pointing to future research directions. Section VII concludes with a summary.

II. Generative Model

In this section, we present the probabilistic model that forms the core of this paper. We use $\{\tilde{I}_n\}$ to denote training images with corresponding label maps $\{\tilde{L}_n\}$, $n = 1, \ldots, N$. We assume the label maps take discrete values from 1 to $\mathcal{L}$ (including a "background" or "unknown" label) at each spatial location. While the training images are defined on a discrete grid, we treat them as spatially continuous functions on $\mathbb{R}^3$ by assuming a suitable interpolator. Let $\Omega \subset \mathbb{R}^3$ be a finite grid where the test subject is defined. We denote $\Phi_n: \Omega \mapsto \mathbb{R}^3$ to be the spatial mapping (warp) from the test subject coordinates to the coordinates of the nth training subject. For simplicity, we assume that $\{\Phi_n\}$ have been precomputed using a pairwise registration procedure, such as the one described in Appendix A. This assumption allows us to shorthand $\{\tilde{I}_n, \Phi_n\}$ and $\{\tilde{L}_n, \Phi_n\}$ with I_n and L_n, respectively, where we drop ~ to indicate that we know the transformation Φ_n that maps the training data into the coordinates of the test subject.

The goal of segmentation is to estimate the label map L associated with the test image I. This can be achieved via maximum-a-posteriori (MAP) estimation

$$\begin{aligned}\widehat{L} &= \underset{L}{\operatorname{argmax}}\; p(L|I;\{\tilde{L}_n, \tilde{I}_n, \Phi_n\}) \\ &= \underset{L}{\operatorname{argmax}}\; p(L, I;\{L_n, I_n\})\end{aligned} \quad (1)$$

where $p(L, I; \{L_n, I_n\})$ denotes the joint probability of the label map and image given the training data.

Instead of using a parametric model for $p(L, I; \{L_n, I_n\})$, we employ a nonparametric estimator, which is an explicit function of the entire training data, not a statistical summary of it, as shown in Fig. 1. The model assumes that the test subject is generated from one or more training subjects, the index or indices of which are unknown. This modeling strategy is parallel to Parzen window density estimators, where the density estimate can be viewed as a mixture distribution over the entire training data, and each new sample is associated with a single training sample, the index of which is unknown and thus is marginalized over. In dealing with images, we may want to allow for this membership index to vary spatially. Therefore we introduce $M: \Omega \mapsto \{1, \ldots N\}$ to denote the latent random field that specifies for each voxel in the test image I, the (membership) index of the training image I_n it was generated from.

In the following, we make the assumption that the image intensity values $I(x)$ and labels $L(x)$ at each voxel are conditionally independent (as illustrated with a plate around these variables in Fig. 1), given the random field M, and the training data $\{L_n, I_n\}$. Furthermore,

we assume that each voxel is generated from a single training subject indexed with $M(x)$, i.e., $p(L(x)|M; \{L_n\}) = p(L(x)|M(x); L_{M(x)})$ and $p(I(x)|M; \{I_n\}) = p\{I(x)|M(x); I_{M(x)})$, which we will shorthand with $p_{M(x)}(L(x); L_{M(x)})$ and $p_{M(x)}(I(x); I_{M(x)})$, respectively. We can thus construct the conditional probability of generating the test image and label map

$$p(L, I|M;\{L_n, I_n\}) = \prod_{x\in\Omega} p(L(x), I(x)|M(x);\{L_n, I_n\}) \quad (2)$$

$$= \prod_{x\in\Omega} p_{M(x)}\left(L(x);L_{M(x)}\right) p_{M(x)}\left(I(x);I_{M(x)}\right). \quad (3)$$

Given a prior on M, we can view the image I and label map L as generated from a mixture model

$$p(L, I;\{L_n, I_n\}) = \sum_{M} p(M) p(L, I|M;\{L_n, I_n\}) \quad (4)$$

where Σ_M denotes the marginalization over the unknown random field M. Substituting (3) into (4) yields the final cost function

$$\widehat{L} = \underset{L}{\operatorname{argmax}} \sum_{M} p(M) \prod_{x\in\Omega} p_{M(x)}(L(x);L_{M(x)}) \times p_{M(x)}(I(x);I_{M(x)}). \quad (5)$$

The conditional independence assumption between the label map and image may seem simplistic at first. Yet conditional independence does not imply independence and the relationship between L and I is given by marginalizing over the unknown M as in (4). Therefore, our model implicitly includes complex dependencies between labels and intensity values. For instance $p(I|L)$, a term commonly modeled explicitly in the segmentation literature can be expressed as

$$\begin{aligned} p(I|L;\{L_n, I_n\}) &= \sum_{M} p(M|L;\{L_n, I_n\}) p(I|L, M;\{L_n, I_n\}) \\ &= \sum_{M} p(M|L;\{L_n\}) p(I|M;\{I_n\}). \end{aligned}$$

Thus, given a model instantiation, the conditional intensity distribution of a particular label at a location of interest can be estimated by examining the training subjects that exhibit that label in the proximity of the location of interest. This is exactly what atlas-based segmentation algorithms do, which underscores the similarity between the proposed probabilistic model and parametric models used in the literature. But unlike atlas-based methods that use a parametric model for $p(I|L)$, the proposed framework explicitly employs the entire training data set.

III. Model Instantiation

This section presents the specific instantiations of the individual terms in (5) that we use in this work to derive segmentation algorithms.

A. Image Likelihood

We adopt a Gaussian distribution with a stationary variance σ^2 as the image likelihood term

$$p_n(I(x);I_n)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}\left(I(x)-\tilde{I}_n(\Phi_n(x))\right)^2\right]. \quad (6)$$

For $\sigma\to\infty$, (6) reduces to an improper distribution $p_n(I(x); I_n) \propto C$, where C is a constant. As we discuss in Section IV-B, this simple model leads to the Majority Voting strategy in label fusion, whereas for a finite σ^2, (6) yields a weighted averaging strategy.

B. Label Prior

In this work, we investigate two representations to define the label prior term $p_n(L(x); L_n)$. One representation uses the logarithm of odds (LogOdds) model based on the signed distance transform [49]. Let $\tilde{D}_n^l$ denote the signed distance transform of label l in training subject n (in the native coordinates), which is assumed to be positive inside the structure of interest. We define the label prior as

$$p_n(L(x)=l;L_n)=\frac{1}{Z_{n,\rho}(x)}\exp\left(\rho\tilde{D}_n^l(\Phi_n(x))\right) \quad (7)$$

where $\rho > 0$ is the slope constant, $Z_{n,\rho}(x)=\sum_{l=1}^{\mathcal{L}}\exp(\rho\tilde{D}_n^l(\Phi_n(x)))$ is the partition function, and $\mathcal{L}$ is the total number of labels including the background label. The prior $p_n(L(x) = l; L_n)$ encodes the conditional probability of observing label l at voxel $x \in \Omega$ of the test image, given that it was generated from the nth training image.

The second representation, commonly used in the literature, employs the probability vector image of $\tilde{L}_n(x)$: each voxel is a length- $\mathcal{L}$ binary vector with the lth entry equal to 1 if $\tilde{L}_n(x) = 1$ and 0 otherwise. To define the label prior $p_n(\cdot; L_n)$, the transformation Φ_n is applied to the probability vector image of $\tilde{L}_n(x)$. In this method, non-grid locations need to be interpolated using a suitable method (e.g., trilinear or nearest neighbor) that ensures positive and normalized probability values. In general, it is well known that trilinear interpolation yields better segmentation results than nearest neighbor interpolation [51], [55]. The LogOdds model of (7) has the advantage of yielding nonzero probabilities everywhere, which makes the use of the logarithm of the probability numerically more stable. As discussed in our experiments presented in Section V-A, we find that the LogOdds model produces more accurate results.

C. Membership Prior

The latent random field $M: \Omega \mapsto \{1, \ldots, N\}$ encodes the local association between the test image and training data. We employ a Markov random field (MRF) prior on M

$$p(M)=\frac{1}{Z_\beta}\prod_{x\in\Omega}\exp\left(\beta\sum_{y\in\mathcal{N}_x}\delta(M(x),M(y))\right) \quad (8)$$

where $\beta \geq 0$ is a scalar parameter, $\mathcal{N}_x$ is a spatial neighborhood of voxel x, Z_β, is the partition function that only depends on β, and δ is the Kronecker delta. This particular type of MRF is often referred to as the Potts model. In our implementation, $\mathcal{N}_x$ includes the immediate six neighbors of each voxel. Similar models have been used in the segmentation literature [64], [72], mainly as priors on label maps to encourage the spatial relationships of labels observed in the training data. In contrast, we use the MRF prior to (indirectly) pool local intensity information from within a neighborhood in determining the association between the test

subject and the training data. Here we adopt a simple form of the MRF that does not include singleton and/or spatially varying terms. This is unlike the common usage of MRFs in the segmentation literature where the label prior typically varies spatially.

The parameter β influences the average size of the local patches of the test subject that are generated from a particular training subject. In this work, we consider three settings of the parameter β. With $\beta = 0$, the model assumes that each test image voxel is generated from the training subjects with equal probability and that the membership is voxel-wise independent. $\beta \rightarrow +\infty$ forces the membership of all voxels to be the same and corresponds to assuming that the whole test subject is generated from a single unknown training subject, drawn from a uniform prior. A positive, finite β encourages local patches of voxels to have the same membership.

IV. Label Fusion Based Image Segmentation

In this section, we derive several label fusion style image segmentation algorithms based on the model and MAP formulation described above. These algorithms correspond to variations in the image likelihood, label prior and membership prior models described in Section III.

A. Local Weighted Voting

Let us assume $\beta = 0$, which, thanks to the adopted simple MRF form, implies that $M(x)$ is independent and identically distributed according to a uniform distribution over all labels for all $x \in \Omega$

$$p(M) = \frac{1}{N^{|\Omega|}} \quad (9)$$

where $|\Omega|$ is the cardinality of the image domain (the number of voxels). Using the image likelihood term of (6), the segmentation problem in (5) reduces to

$$\widehat{L}(x) = \underset{l \in \{1,\ldots,\mathscr{L}\}}{\operatorname{argmax}} \sum_{n=1}^{N} p_n(L(x)=l; L_n)\, p_n(I(x); I_n). \quad (10)$$

This optimization problem can be solved by simply comparing $\mathscr{L}$ numbers at each voxel: the fused label of each voxel is computed via a local weighted (fuzzy) voting strategy. The local image likelihood terms serve as weights and the label prior values serve as votes. Therefore, at each voxel, training images that are more similar to the test image at the voxel *after* registration are weighted more. Interestingly, a similar approach was recently proposed in the context of CT cardiac segmentation by Isgum *et al.* [32] where the transferred training labels are fused in a weighted fashion. The heuristic weights proposed in that paper have a different form however and are spatially smoothed with a Gaussian filter to pool local neighborhood information. In Section IV-D, we discuss a more principled approach to aggregate statistical information from neighboring voxels into the weighted label fusion procedure.

B. Majority Voting

Majority voting, which has been widely used as a label fusion method [30], [51] can be derived as a special case of *Local Weighted Voting*. The key modeling assumption is to set $\sigma \rightarrow \infty$ in the image likelihood term, effectively using an improper distribution $p_n(I(x); I_n) \propto C$ and assigning equal weight to all training subjects, which reduces (10) to

$$\widehat{L}(x)=\underset{l\in\{1,\ldots,\mathscr{L}\}}{\operatorname{argmax}}\sum_{n=1}^{N}p_n(L(x)=l;L_n). \quad (11)$$

If we use the probability vector image of $\tilde{L}_n(x)$ to define the label prior, we arrive at the traditional Majority Voting algorithm where each training image casts a single, unit vote, with no regards to the similarity between the training image and the test image. If one uses nearest neighbor interpolation, each vote corresponds to one particular label [30], whereas tri-linear interpolation yields a fuzzy voting strategy with each vote potentially spread over multiple labels [51].

C. Global Weighted Fusion

Here, we consider $\beta \rightarrow +\infty$. As we now show, this results in an algorithm where, at each voxel, training images that are *globally* more similar to the test image *after* registration are weighted more. With $\beta \rightarrow +\infty$, the membership prior defined in (8) only takes nonzero values if membership values at all voxels are equal, i.e.,

$$p(M)=\begin{cases}\frac{1}{N}, & \text{if } M(x)=j\ \forall x\in\Omega,\ \exists j\\ 0, & \text{otherwise.}\end{cases} \quad (12)$$

Thus, (4) is equivalent to a mixture model where the test subject is assumed to be generated from a single, unknown training subject

$$p(L,I;\{L_n,I_n\})=\frac{1}{N}\sum_{n=1}^{N}p_n(L;L_n)p_n(I;I_n). \quad (13)$$

The segmentation problem in (5) reduces to

$$\widehat{L}(x)=\underset{L}{\operatorname{argmax}}\sum_{n=1}^{N}\prod_{x\in\Omega}p_n(L(x);L_n)p_n(I(x);I_n). \quad (14)$$

Equation (14) cannot be solved in closed form. However, an efficient solution to this MAP formulation can be obtained via expectation maximization (EM) [20]. Appendix B contains the derivations of the algorithm. Here, we present the summary.

1. *E-Step:* Let $\hat{L}^{(i-1)}(x)$ denote the segmentation estimate in the $(i-1)$th iteration. The E-step updates the posterior of the membership associated with each training image

$$m_n^{(i)} \propto \prod_{x\in\Omega}p_n(I(x);I_n)p_n(\widehat{L}^{(i-1)}(x);L_n) \quad (15)$$

where $m_n^{(i)}$ is the current estimate of the posterior probability that the test image was generated from the nth training image. Therefore $\sum_n m_n^{(i)}=1$. The E-step in (15) determines a single, global membership for each training image, based on all voxels.

2. *M-Step:* The M-step updates the segmentation estimate (16)

$$\widehat{L}^{(i)}(x)=\underset{l\in\{1,\ldots,\mathcal{L}\}}{\operatorname{argmax}}\sum_{n=1}^{N} m_n^{(i)}\log p_n(L(x)=l;L_n) \quad (16)$$

$$=\underset{l\in\{1,\ldots,\mathcal{L}\}}{\operatorname{argmax}}\sum_{n=1}^{N} m_n^{(i)}\tilde{D}_n^{l}(\Phi_n(x)) \quad (17)$$

where $\tilde{D}_n^l$ denotes the signed distance transform of label l in training subject n. We note that (17) uses the LogOdds model of (7). The M-step in (16) performs an independent optimization at each voxel $x \in \Omega$; it determines the mode of a length-$\mathcal{L}$ vector, where $\mathcal{L}$ is the number of labels. This vector is computed as *a weighted average of log label priors* (i.e., signed distance transforms, $\tilde{D}_n^l$). The global memberships computed in the previous E-step serve as weights and depend on the current label map estimate. The algorithm iterates between the weighted averaging of the M-step in (16) and the global weight computation of the E-step in (15). We initialize the EM algorithm with $m_n^{(1)} \propto \prod_{x\in\Omega} p_n(I(x);I_n)$ and terminate the algorithm when the average change in the membership weights is less than a small threshold, e.g., 0.01. We found that our EM algorithm typically converges in fewer than 10 iterations.

D. Semi-Local Weighted Fusion

Finally, we consider a finite, positive β that results in an algorithm, where at each voxel, training images that are more similar to the test image in a local neighborhood of the voxel *after* registration are weighted more. Because the MRF prior couples neighboring voxels, the exact marginalization in (5) becomes computationally intractable.

An efficient *approximate* solution can be obtained via variational EM (also referred to as variational mean field) [33]. Variational EM uses a fully factorized distribution q over the random field M

$$q(M)=\prod_{x\in\Omega} q_x(M(x)) \quad (18)$$

to approximate the posterior distribution $p(M|\hat{L}, I; \{L_n, I_n\})$ of the random field M given the segmentation estimate $\hat{L}$, test image and the training data. $q_x(n)$ can thus be viewed as the approximate posterior probability that voxel x was generated from the n-th training image. The algorithm alternates between updating the approximate posterior q (the E-step) and the segmentation $\hat{L}$ (M-step). Appendix C includes the details of the derivation. Here, we present the summary.

1. *E-Step:* Let $\hat{L}^{(i-1)}(x)$ denote the segmentation estimate in the $(i-1)$th iteration. The approximate posterior q is the solution of the following fixed-point equation:

$$q_x^{(i)}(M(x)) \propto p_{M(x)}\left(I(x);I_{M(x)}\right) \times p_{M(x)}\left(\widehat{L}^{(i-1)}(x);L_{M(x)}\right)\exp\left(\beta\sum_{y\in\mathcal{N}_x} q_y^{(i)}(M(x))\right) \quad (19)$$

where $\sum_n q_x^{(i)}(n)=1$. We compute (19) repeatedly, to iteratively solve for $q^{(i)}$.

2. *M-Step:* The M-step updates the segmentation estimate $\hat{L}$ as follows:

$$\widehat{L}^{(i)}(x)=\underset{l\in\{1,\ldots,\mathcal{L}\}}{\operatorname{argmax}}\sum_{n=1}^{N}q_x^{(i)}(n)\log p_n(L(x)=l;L_n) \quad (20)$$

$$=\underset{l\in\{1,\ldots,\mathcal{L}\}}{\operatorname{argmax}}\sum_{n=1}^{N}q_x^{(i)}(n)\tilde{D}_n^l(\Phi_n(x)). \quad (21)$$

Equation (21) uses the LogOdds model of (7) and is computed independently for each voxel. Optimizing (21) entails determining the mode of a length- $\mathcal{L}$ vector, which is a *weighted average of the log label priors* corresponding to each training subject. The current approximate posterior $q^{(i)}$ for each training subject serve as weights.

The variational EM algorithm consists of two levels of iterations: the inner loop that repeatedly computes (19) in the E-step and the outer loop that alternates between the E- and M-steps, until convergence. In the inner loop, at each iteration all q_x's are updated using (19) and the neighboring values $\{q_y : y \in \mathcal{N}_x\}$ from the previous iteration. Once this inner loop converges, the algorithm updates the segmentation using (21). To determine the convergence of the outer loop, one can monitor the change in the segmentation estimate. In practice, we terminate the algorithm when less than a predetermined fraction, e.g., 0.01% of voxels change their segmentation estimate from one iteration to the next. Typically convergence is achieved in fewer than 10 iterations.

V. Experiments

In this section, we present two sets of experiments. In the first experiment we compare automatic segmentation results against manual delineations to objectively quantify the accuracy of segmentation algorithms. The second experiment employs a separate collection of brain MRI scans from 282 subjects to demonstrate that hippocampal volume measurements obtained using the proposed label fusion framework can detect subtle volume changes associated with aging and Alzheimer's disease.

A. Experiment I: Comparison Against Manual Segmentation

The first set of experiments employs 39 brain MRI scans and corresponding manual delineations of nine anatomical regions of interest (ROI) in two hemispheres. The images were selected from a large data set, including an Alzheimer's cohort, the recruitment of which is described elsewhere [19], [35], [38]. The 39 subjects were selected to span a wide age range and reflect a substantial anatomical variation due to dementia pathology. We note that these are the same subjects used to construct FreeSurfer's released probabilistic segmentation atlas. Out of the 39 subjects, 28 were healthy and 11 were patients with questionable (N = 5, Clinical Dementia Rating 0.5) or probable Alzheimer's (N = 6, CDR 1). Ten of the healthy subjects were young (less than 30 years), nine middle-aged (between 30 and 60 years), and nine old (older than 60 years). The MRI images are of dimensions 256 × 256 × 256, 1 mm isotropic voxels and were computed by averaging three or four scans. Each scan was a T1-weighted MP-RAGE, acquired on a 1.5 T Siemens Vision scanner. All scans were obtained in a single session. Acquisition details are as follows: TR 9.7 ms, TE 4.0 ms, TI 20 ms, Flip angle 10°. These high quality images were then gain-field corrected and skull-stripped. All the preprocessing steps were carried out using FreeSurfer tools [69]. The anatomical ROIs we used[1] are white matter (WM), cerebral cortex (CT), lateral ventricle (LV), hippocampus (HP), thalamus (TH), caudate (CA), putamen (PU), pallidum

(PA), and amygdala (AM). The labeling protocol we employed was developed by the Center for Morphometric Analysis and has been published and validated elsewhere [13], [28], [37], [59]. An example segmentation obtained via the local weighted voting method of Section IV-A is visualized in Fig. 2.

We use a volume overlap measure known as the Dice score [21] to quantify the quality of automatic segmentations. Given an automatic segmentation $\hat{L}$ and the corresponding manual segmentation L, the Dice score of label l is defined as

$$\mathrm{Dice}(l;\widehat{L},L)=2\frac{|\{x\in\Omega|L(x)=l\&\widehat{L}(x)=l\}|}{|\{x\in\Omega|L(x)=l\}|+|\{x\in\Omega|\widehat{L}(x)=l\}|} \quad (22)$$

where $|\cdot|$ denotes set cardinality. The Dice score varies between 0 and 1, with 1 indicating a perfect agreement between the two segmentations.

1. *Setting the Free Parameters Through Training:* The proposed label fusion algorithms have two stages, registration and label fusion, each with several input parameters. To set these parameters properly, we initially performed *training* on nine out of the 39 subjects. These nine subjects were then only used as training subjects during the testing phase. Therefore, all results reported in this paper are on the remaining 30 subjects and reflect generalization performance.

   The registration stage has two independent parameters (as described in Appendix A): γ controls the step size in the Gauss–Newton optimization and α determines the smoothness of the final warp. We registered 20 random pairs of the nine training subjects for a range of values of γ and α. For each pair of subjects, we measured pairwise overlap by computing the Dice score between the warped manual labels of the "moving" subject and the manual labels of the "fixed" subject. We then selected (γ^*, α^*) that resulted in the best registration quality as measured by the average pairwise label overlap.

   The label fusion stage also has several independent parameters, depending on the method used. These include the standard deviation σ of the image likelihood in (6), the slope ρ of the distance transform used to compute the label prior in (7), and the Markov weight β which is nonzero for the semi-local method in Section IV-D and controls the average size of the image patches associated with the same training subject.

   To determine ρ, we performed nine leave-one-out segmentations on the training subjects using the Majority Voting method of Section IV-B and label prior model of (7) for a range of ρ values. The value that achieved the best segmentation accuracy was $\rho^* = 1$. We employed Local Weighted Voting (Section IV-A) and Global Weighted Fusion (Section IV-C) to determine a local and global optimal value for σ^* (10 and 30), respectively. The optimal standard deviation for σ^* the local model was then used to determine the optimal value for β^* (0.75) for the semi-local model.

   We performed leave-one-out cross-validation on the 30 test subjects using these optimal parameters. For each test subject, all remaining 38 subjects were treated as training subjects.

[1]The data included more ROIs, e.g., cerebellum, brainstem, third ventricle, which we did not use in our analysis.

2. *Comparison of Label Prior Models:* Using the Majority Voting method (Section IV-B), we compare three different label prior models (Section III-B): the LogOdds (based on the signed distance transform) model of (7) and two instantiations of the common approach that interpolates the vector image $\tilde{L}_n(x)$ of indicator probability vectors, based on nearest neighbor interpolation (e.g., [30]) or tri-linear interpolation (e.g., [51]). Fig. 3 shows a box-plot of Dice scores for these three different models and all the ROIs. These results indicate that the LogOdds representation provides a significantly better label prior for the label fusion framework. This finding is in agreement with the main point of [52], where the authors propose to employ signed distance transforms when "fusing" labels, essentially arguing that this representation is more suitable for averaging complex shapes such as the cortex. In the remainder of the paper, we use the LogOdds model based on the signed distance transform to compute the label prior. Note that the corresponding *majority voting* procedure is different from simply averaging the signed distance transforms as proposed in [52] since the signed distance transforms are exponentiated [see (7) and (11)] and converted into probabilities before averaging. Interestingly, however, in Global and Semi-local Weighted Fusion, the algorithms apply a weighted averaging to the signed distance transforms directly at each iteration [see (17) and (21)].

3. *Comparison of Label Fusion Methods and Benchmarks:* In this section we provide a comparison between the three weighted label fusion algorithms we derived in our framework and four benchmarks.

   The first benchmark is the whole-brain segmentation tool available in the FreeSurfer software package [69]. The FreeSurfer segmentation tool employs a unified registration-segmentation procedure that models across-scanner intensity variation [24], [29]. We consider this a state-of-the-art whole-brain segmentation tool since numerous imaging studies across multiple centers have shown FreeSurfer's robustness and accuracy as a segmentation tool. Furthermore, this is our only benchmark that does not rely on the preprocessing step of pairwise registering each training image with the test image. FreeSurfer's segmentation tool uses a probabilistic atlas constructed from the training data. In our experiments, we constructed a separate leave-one-out atlas for each test subject based on the remaining 38 subjects.

   The second benchmark is the Majority Voting scheme based on the LogOdds prior, which is similar to the shape averaging method proposed in [52] and other voting based algorithms, e.g., [30], [51].

   Our third benchmark uses the STAPLE algorithm [67] to fuse the propagated labels. STAPLE was originally developed to combine manual tracings of the same subject from multiple raters and estimate the underlying "ground truth" labeling. In doing so, it ignores MRI intensity information and utilizes a probabilistic model that encodes and estimates the rater performance for each label. STAPLE employs Expectation Maximization to efficiently solve the estimation problem. One can also use STAPLE to combine the transferred labels from each training subject, as suggested in [53]. In our experiments we implemented a multilabel version of STAPLE to combine the transferred training labels.

   Our fourth benchmark is a modified version of majority voting that efficiently selects a subset of training subjects that are "closest" to the test subject to vote on the labels [1]. Various strategies to define the similarity between images have been proposed. In our experiments, we found that the following strategy gave us the best results. First, all training subjects were co-registered using an affine transformation

model. The test subject was then normalized to this space by an affine registration with the group mean image. Next, we computed the sum of squared intensity differences (SSD) between each training subject and test subject within a predefined mask. Similar to [1], the mask was computed as the intersection of the foreground labels in all training subjects. Finally, the training subjects that had the smallest SSD were used for majority voting. In the results we present here, we fix the number of training subjects that were used to 10 and call the algorithm "Majority10." Later in this section, we investigate the effects of varying the number of training subjects.

Fig. 4 reports segmentation accuracy for benchmarks and the three weighted label fusion methods: Local Weighted Voting, Global Weighted Fusion, and Semi-Local Weighted Fusion. Table I provides the mean Dice scores averaged over all subjects and both hemispheres. Fig. 5 provides an overall comparison between the average Dice scores achieved by the algorithms.

Semi-local Weighted Fusion yields the most accurate segmentations in all ROIs but the Cortex (CT). The difference with the benchmarks is statistically significant ($p < 0.05$, Bonferroni corrected) for all ROIs, except the CT for which FreeSurfer yields the best accuracy. Similarly, Local Weighted Voting yields statistically better segmentations than the benchmarks.

On average, Local Weighted Voting and Semi-local Weighted Fusion yield better segmentations than Global Weighted Fusion, mainly due to the large improvement in the white matter, cerebral cortex and lateral ventricles, the segmentation of which clearly benefits from the additional use of local intensity information. A paired permutation test between the Local and Semi-local models reveals that in all ROIs, a statistically significant improvement is achieved with the MRF model that pools local intensity information ($p < 0.05$, Bonferroni corrected). Yet, as can be seen from Fig. 6, this improvement is overall quite modest: less than 0.14% per ROI.

Majority Voting which has gained recent popularity [1], [30], [42], [51], performs significantly worse than the weighted label fusion methods. This result highlights the importance of incorporating image similarity information into the label fusion framework. We note, however, that the results we report for our Majority Voting implementation are lower than the ones reported in [30]. This might be due to differences in the data and/or registration algorithm. Specifically, normalized mutual information (NMI) was used as the registration cost function in [30]. Entropy-based measures such as NMI are known to yield more robust alignment results. We leave a careful analysis of this issue to future work.

Majority10 performs slightly better than Majority Voting. The improvement is particularly significant in subcortical ROIs such as the caudate. STAPLE, an alternative weighted fusion strategy, also yields slightly better average segmentation accuracy than Majority Voting. STAPLE's performance, however, is significantly worse than the three weighted label fusion algorithms derived based on the proposed probabilistic framework. Once again, this difference underscores the importance of employing the MRI intensity information in determining the weights for label fusion. FreeSurfer, which we consider to represent the state-of-the art atlas based segmentation, on average, yields better segmentation accuracy than our remaining benchmarks. Yet we stress that FreeSurfer integrates registration and segmentation, while the performance of the remaining benchmarks were limited by our choice of the pairwise registration preprocessing step.

Fig. 7 illustrates the most common mistakes made by Global Weighted and Semi-local Weighted Fusion. Overall, we observe that Global Weighted Fusion tends to over-segment convoluted shapes such as the cortex. This is probably due to the difficulty of aligning such shapes. Majority voting has no way of recovering from such errors since it does not utilize image intensity information. Local Weighted Voting and Semi-local Weighted Fusion do not employ information on neighborhood structure, whereas atlas-based methods, such as FreeSurfer, do. For example, FreeSurfer uses a non-stationary, anisotropic Markov Random Field model to encode that the pallidum is more medial to the putamen and there is no white matter in between the two, i.e., they border each other. Our label fusion framework does not model such high-level topological information and thus may yield segmentations where anatomical relationships are violated.

4. *The Effect of the Number of Training Subjects:* In the previous section, for all the algorithms we employed a leave-one-out validation strategy, where for each test subject all remaining 38 subjects were treated as the training data. In this section, we explore how the accuracy results vary as one varies the number of training subjects. This point has received considerable attention in prior work, e.g., [1], [30]. We investigate two strategies: 1) randomly selecting a set of training subjects, 2) selecting the best training subjects that are globally most similar to the test subject. As described in the previous section, based on [1], [42], we implemented strategy 2 by computing the sum of squared differences between the intensity values of the test subject and each training subject in a predefined mask after affine registration. This strategy does not add a substantial computational burden to the algorithm, since the training subjects are co-registered offline and the affine normalization of the test subject takes negligible time compared with the pairwise registrations required for the label fusion stage. We compare the performance of two segmentation algorithms using these two strategies: Majority Voting and Local Weighted Voting. Fig. 8 reports the mean Dice score as a function of the number of training subjects. The mean Dice score is an average over all ROIs, all 30 test subjects, and at least 20 random trials (in strategy 1). The individual results for each ROI are qualitatively similar to this average plot and thus are not shown. For strategy 1, i.e., the random selection of the training subjects, average segmentation performance monotonically improves as a function of the number of training subjects for both Majority Voting and Local Weighted Voting. With Majority Voting, this monotonic behavior was also observed and modeled in [30]. The difference between the two label fusion algorithms is approximately constant. The rate of improvement seems to be flattening around mid-thirties, suggesting that we should not expect a significant improvement in performance with the addition of more subjects within the observed anatomical variation. For a fixed number of training subjects, selecting the globally most similar training subjects, i.e., strategy 2 improves the performance of both algorithms. With strategy 2, Majority Voting's performance however is not monotonic and starts to decrease beyond 10. In agreement with the empirical observations in [1], as the number of training subjects approaches the total number of available training subjects in the database, the performance of the strategies for training set selection converges to the same level, the accuracy obtained by employing the whole dataset for training. This level of accuracy is higher for Local Weighted Voting, the performance of which monotonically increases with the number of training subjects for both strategies. Interestingly, with strategy 2, the performance of Local Weighted Voting flattens around 15 training subjects. Based on this observation, we conclude that we can use strategy 2 to speed up Local Weighted Voting substantially with minimal effect on segmentation accuracy.

5. *The Effect of the MRF Prior:* To investigate the effect of the MRF membership prior we applied the Semi-local Weighted Fusion method with four different values of $\beta = \{0.5, 0.75, 1.0, 1.25\}$ to the 30 test subjects. Note that during the training phase, we established $\beta = 0.75$ as optimal. Fig. 9 reports the average Dice scores for Semi-local Weighted Fusion with these β values, Global Weighted Fusion, which corresponds to $\beta \rightarrow \infty$, and Local Weighted Voting, which corresponds to $\beta = 0$.

   In four out of the nine ROIs (WM, CT, LV, TH) Semi-local Weighted Fusion with $\beta = 0.075$ achieves the best results. In the remaining ROIs, $\beta = 0.75$ yields one of the best. This underscores the importance of pooling local intensity information in the label fusion procedure and the success of our training phase. The improvement over Global Weighted Fusion is particularly pronounced for the white matter, cerebral cortex and lateral ventricle, which are essentially defined by tissue intensity. For putamen (PU) and pallidum (PA), Global Weighted Fusion achieves one of the best results. This suggests that the segmentation of these two ROIs has little to gain by employing local intensity information. The gap between the Semi-local ($\beta = 1.25$) and Global Weighted Fusion in these two ROIs is probably because variational EM fails to find good optima for these regions whereas Expectation Maximization is better at exploring the space of solutions in the simpler case of global memberships. Furthermore, this difference can be due to the fact that the image intensities in the putamen and pallidum are similar to neighboring structures and thus local intensity information is less useful in segmenting these ROIs.

6. *Runtime:* Table II lists the average run-times for the seven algorithms compared above. Majority10 and FreeSurfer are the fastest algorithms with less than 10 h of CPU time required for each test subject. Majority10 uses only 10 training subjects, which are globally most similar to the test subject as measured by the sum of squared differences after affine-normalization. The initial training subject selection stage takes about an hour. The second stage, i.e., Majority Voting with 10 training subjects takes about a quarter of what Majority Voting takes with all 38 training subjects. FreeSurfer, on the other hand, employs a parametric atlas and needs to compute only a single registration. The remaining algorithms take more than 20 h of CPU time on a modern machine (Intel Xeon 3 GHz with a 32 GB RAM), most of which is dedicated to the many registrations of the test image with the training data. Local Weighted Voting and Majority Voting require minimal computation time once the registrations are complete, since they simply perform voxelwise averaging. The remaining three algorithms (STAPLE, Global and Semi-local Weighted Fusion) employ iterative optimization methods (EM, EM and variational EM, respectively) and require longer run-times. It is important to note that these run times can be reduced substantially using the same preselection strategy as Majority 10. In particular, our experiments with Local Weighted Voting suggest that we can lower the run time of this method by at least a half with almost no reduction in accuracy.

B. Experiment II: Hippocampal Volumetry

In a second set of experiments, we aim to demonstrate that the proposed label fusion framework yields accurate volumetric measurements of the hippocampus. Hippocampal volume has been shown to correlate with aging and predict the onset of probable Alzheimer's Disease [25], [34].

We use a FreeSurfer tool (mri_label_volume) [69] that computes volumetric measurements from a given label map to obtain hippocampal volume measurements. This tool uses a

partial volume model for boundary voxels to accurately estimate the volume of a particular ROI. As an initial experiment, we compare hippocampal volume measurements obtained via automatic segmentations (FreeSurfer, Global Weighted Fusion, Local Weighted Voting, and Semi-local Weighted Fusion) against the ones computed from manual delineations in the 39 subjects used in the first experiment. Fig. 10(a) reports volume differences between the automatic and manual segmentations;

Fig. 10(b) shows relative volume difference values, defined as in [16]

$$\text{Volume Difference}(l;\widehat{L}, L)=2\frac{|V(l,\widehat{L})-V(l,L)|}{V(l,\widehat{L})+V(l,L)} \quad (23)$$

where $V(l, L)$ denotes the computed volume of label in label l map L. These results indicate that both Local Weighted Voting and Semi-local Weighted Fusion provide more accurate hippocampal volume measurements than Global Weighted Fusion and Majority Voting.

Since Local Weighted Fusion is much less computationally expensive than Semi-local Weighted Fusion, we choose to employ the former algorithm in the second part of the experiment, where we automatically segment brain MRI scans of 282 individuals. This data set excludes the 39 subjects used in training. The MRI images are of dimensions 256 × 256 × 256 with 1 mm isotropic voxels and were obtained using the same processing and acquisition protocol as the 39 images of the first experiment. The population included young (less than 30 years, $N = 105$), middle-aged (between 30 and 60 years, $N = 30$) and old subjects (older than 60 years $N = 78$) (see Fig. 11), in addition to patients suffering from very mild to mild Alzheimer's Disease (AD) ($N = 69$). Among the patients, 48 subjects met the Clinical Dementia Rating (CDR) [44] criteria of questionable AD (CDR 0.5) and 21 subjects had probable AD (CDR 1). The CDR 0.5 subpopulation contained 32 women, whereas the probable AD group included 21 women. The groups did not differ in years of education. Detailed descriptions of the recruitment procedures and criteria for subject recruitment have been published elsewhere [19], [35], [38].

Based on age and clinical data, we subdivided the 282 subjects into five groups. The first three groups contained healthy individuals of different age groups: young, middle-aged, and old. The fourth group included the questionable AD patients (CDR 0.5) and the fifth group included the probable AD (CDR 1.0) subpopulation.

Fig. 12 shows the average hippocampal volume measurements for these five groups. Volumetric reduction due to aging and AD can be seen from this figure. These findings are in agreement with known hippocampal volume changes in AD and aging [23] and demonstrate the use of the proposed label fusion method on a large pool of subjects, for which manual segmentation may not be practical.

VI. Discussion

Our experiments demonstrate the accuracy and usefulness of the label fusion framework as a segmentation tool. The proposed framework yields better accuracy than current state-of-the-art atlas-based segmentation algorithms, such as FreeSurfer.

The proposed framework should be viewed as an initial attempt to generalize segmentation algorithms based on label fusion, or a multi-atlas approach, which have recently shown promise and gained popularity with hardware advancements and developments of fast registration algorithms. In this paper, we investigated several modeling assumptions and

derived four different instantiations of label fusion, one of which is the popular Majority Voting.

Majority Voting simply determines the most frequent label at each voxel, where each training subject gets an equal vote. Yet, recent work suggests that incorporating the similarity between the test image and training subjects can improve segmentation quality. For example, [1] employs a subset of training subjects that are close in age to the test subject. Alternative strategies include using an image-based measure to quantify anatomical similarity, either at a local or global level. This similarity can then weigh the label votes during fusion, where more similar training subjects are given a larger weight.

Our theoretical development based on the proposed nonparametric probabilistic model yields three such algorithms, which solve the same problem for different settings of a single model parameter β. This parameter controls the interactions between neighboring voxels in the Markov prior we construct on the latent membership random field encoding the (unknown) association between the test subject and training data. Smaller β values allow for this association to vary more locally. Specifically, $\beta = 0$ treats each voxel independently, whereas $\beta \rightarrow \infty$ corresponds to assuming a single association for the whole image. A finite, nonzero β encourages local patches of voxels to have the same membership.

These three cases are solved with different inference algorithms. The most efficient case corresponds to $\beta = 0$, where the global optimum can be computed via simple voxelwise counting. The other two cases are solved with more expensive iterative optimization methods, such as Expectation Maximization and Variational EM. Exact inference for the finite, nonzero β case is intractable, yet our experiments suggest that approximate numerical solutions yield good segmentation accuracy.

The development of the proposed framework makes several simplifying assumptions. In the following, we discuss a number of directions that can be explored to relax these assumptions. We consider these as important avenues for future research, which promise to improve the performance of label fusion style segmentation.

1. In the graphical model of Fig. 1, we made the convenient assumption that the transformations $\{\Phi_n\}$ are known and solved for these in a preprocessing pairwise registration step (see Appendix A). Ideally, however, one would like to integrate over all possible transformations, which has a prohibitively high computational cost. Recent work attempted to approximate this integration for a *single* registration [3], [40]. A more practical approach is to compute the registrations jointly with the segmentations, cf. [8], [72]. Here, we avoided this particular route, since the multiple registrations performed between the test subject and training data were already computationally challenging.
2. The simple additive Gaussian noise model presented in Section III-A has two crucial consequences: 1) the registration cost function is a sum of squared intensity differences, and 2) in weighted label fusion, the weights are a function of sum of squared intensity differences, i.e., anatomical similarity is measured based on squared differences of intensity values. This model makes the algorithm sensitive to intensity variations due to imaging artifacts. Thus, the presented algorithms are only suitable for intensity-normalized images. An alternative strategy is to employ a more sophisticated image likelihood model that would motivate information theoretic similarity measures, such as mutual information.
3. The main drawback of label fusion style algorithms is the computational complexity introduced by the multiple pairwise registrations and the manipulation of the entire training data. Traditional atlas-based segmentation approaches avoid

this problem by using parametric models of anatomical variation in a single coordinate system. In recent work, we used a mixture modeling strategy, called iCluster, to model anatomical heterogeneity with multiple atlases [54]. We believe a combination of the label fusion framework presented in this paper and iCluster can be employed to reduce the computational burden by summarizing the training data with a small number of *templates* [66].

4. An alternative strategy to reduce the computational demand of label fusion is to employ a nonparametric model in a single coordinate system, to which the test subject is normalized with a single registration procedure. This approach, which entails the co-registration of the training subjects akin to atlas-based segmentation, was recently shown to produce accurate segmentation [5], [17]. The application of this strategy within the proposed label fusion framework is a direction to be explored.

5. Another strategy to reduce computational burden is to preselect the most useful training subjects and apply label fusion on these, as recently proposed by Aljabar *et al.* [1]. We explored one particular instantiation of this approach, where the subset of training subjects was selected to include the training subjects globally most similar to the test subject after affine normalization. It is clear that this criterion to preselect the most relevant training subjects is related to our definition of the image likelihood term $p(I; I_n)$. Yet, a crucial difference is that the image likelihood term is computed by nonlinearly registering the training and test images, while the preselection is done based on an affine normalization. Alternative preselection strategies should also be investigated.

VII. Conclusion

In this paper, we investigated a generative model that leads to label fusion style image segmentation methods. Within the proposed framework, we derived several algorithms that combine transferred training labels into a single segmentation estimate. With a dataset of 39 brain MRI scans and corresponding label maps obtained from an expert, we empirically compared these segmentation algorithms with FreeSurfer's widely-used atlas-based segmentation tool [69]. Our results demonstrate that the proposed framework yields accurate and robust segmentation tools that can be employed on large multi-subject datasets. In a second experiment, we employed one of the developed segmentation algorithms to compute hippocampal volumes in MRI scans of 282 subjects. A comparison of these measurements across clinical and age groups indicate that the proposed algorithms are sufficiently sensitive to detect hippocampal volume differences associated with early Alzheimer's Disease and aging.

Acknowledgments

Support for this research is provided in part by: NAMIC (NIH NIBIB NAMIC U54-EB005149), the NAC (NIH NCRR NAC P41-RR13218), the mBIRN (NIH NCRR mBIRN U24-RR021382), the NIH NINDS R01-NS051826 grant, the NSF CAREER 0642971 grant, NCRR (P41-RR14075, R01 RR16594-01A1), the NIBIB (R01 EB001550, R01EB006758), the NINDS (R01 NS052585-01), the MIND Institute, and the Autism and Dyslexia Project funded by the Ellison Medical Foundation. B. T. T. Yeo is funded by the A*STAR, Singapore. K. Van Leemput was supported in part by the Academy of Finland, grant number 133611.

References

1. Aljabar P, Heckemann RA, Hammers A, Hajnal JV, Rueckert D. Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. Neuroimage. 2009; 46(3):726–738. [PubMed: 19245840]

2. Allassonnière S, Amit Y, Trouve A. Towards a coherent statistical framework for dense deformable template estimation. J R Stat Soc B. 2007; 69:3–29.
3. Allassonnière, S.; Kuhn, E.; Trouvé, A. MAP estimation of statistical deformable template via nonlinear mixed effect models: Deterministic and stochastic approaches. Math. Foundations Computat. Anat. (MFCA) Workshop MICCAI 2008 Conf; 2008.
4. Arsigny, V.; Commowick, O.; Pennec, X.; Ayache, N. Proc of MICCAI. Vol. 4190. New York: Springer; 2006. A log-Euclidean framework for statistics on diffeomorphisms; p. 924-931.Lecture Notes Computer Science
5. Artaechevarria, X.; Munoz-Barrutia, A.; de Solorzano, CO. Efficient classifier generation and weighted voting for atlas-based segmentation: Two small steps faster and closer to the combination oracle. SPIE Med. Imag; 2008; 2008. p. 6914
6. Artaechevarria X, Munoz-Barrutia A, de Solorzano CO. Combination strategies in multi-atlas image segmentation: Application to brain MR data. IEEE Trans Med Imag. Aug; 2009 28(8):1266–1277.
7. Ashburner J. A fast diffeomorphic image registration algorithm. Neuroimage. 2007; 38(1):95–113. [PubMed: 17761438]
8. Ashburner J, Friston K. Unified segmentation. Neuroimage. 2005; 26:839–851. [PubMed: 15955494]
9. Avants, BB.; Grossman, M.; Gee, JC. Biomedical Image Registration. Vol. 4057. New York: Springer; 2006. Symmetric diffeomorphic image registration: Evaluating labeling of elderly and neurodegenerative cortex and frontal lobe; p. 50-57.LNCS
10. Awate S, Tasdizen T, Foster N, Whitaker R. Adaptive Markov modeling for mutual-information-based, unsupervised MRI brain-tissue classification. Med Image Anal. 2006; 10(5):726–739. [PubMed: 16919993]
11. Blezek D, Miller J. Atlas stratification. Med Image Anal. 2007; 11(5):443–457. [PubMed: 17765003]
12. Cachier P, Bardinet E, Dormont D, Pennec X, Ayache N. Iconic feature based non-rigid registration: The PASHA algorithm. Comput Vis Image Understand. 2003; 89(2–3):272–298.
13. Caviness VS, Filipek PA, Kennedy DN. Magnetic resonance technology in human brain science: Blueprint for a program based upon morphometry. Brain Develop. 1989; 11:1–13.
14. Christensen GE, Johnson HJ. Consistent image registration. IEEE Trans Med Imag. Jul; 2001 20(7):568–582.
15. Cocosco C, Zijdenbos A, Evans A. A fully automatic and robust brain MRI tissue classification method. Med Image Anal. 2003; 7(4):513–527. [PubMed: 14561555]
16. Collins DL, Holmes CJ, Peters TM, Evans AC. Automatic 3-d model-based neuroanatomical segmentation. Human Brain Mapp. 1995; 3(3):190–208.
17. Commowick, O.; Warfield, S.; Malandain, G. Using Frankensteins creature paradigm to build a patient specific atlas. Proc. of MICCAI; 2009; New York: Springer; 2009. p. 993-100.Lecture Notes Computer Science
18. D'Agostino, E.; Maes, F.; Vandermeulen, D.; Suetens, P. Biomedical Image Registration. Vol. 4057. New York: Springer; 2006. A unified framework for atlas based brain image segmentation and registration; p. 136-143.Lecture Notes Computer Science
19. Daly E, Zaitchik D, Copeland M, Schmahmann J, Gunther J, Albert M. Predicting conversion to alzheimer disease using standardized clinical information. Arch Neurol. 2000; 57:675–680. [PubMed: 10815133]
20. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Series B (Methodological). 1977; 39(1):1–38.
21. Dice LR. Measures of the amount of ecologic association between species. Ecology. 1945; 26(3): 297–302.
22. Fischl B, Busa NRE, Augustinack J, Hinds O, Yeo BTT, Mohlberg H, Amunts K, Zilles K. Cortical folding patterns and predicting cytoarchitecture. Cerebral Cortex. 2008; 18(8):1973–1980. [PubMed: 18079129]
23. 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. Whole brain segmentation:

Automated labeling of neuroanatomical structures in the human brain. Neuron. 2002; 33(3):341–355. [PubMed: 11832223]

24. Fischl B, Salat DH, van der Kouwe AJW, Makris N, Ségonne F, Quinn BT, Dale AM. Sequence-independent segmentation of magnetic resonance images. Neuroimage. 2004; 23:69–84.
25. Fox NC, Warrington EK, Freeborough PA, Hartikainen P, Kennedy AM, Stevens JM, Rossor MN. Presymptomatic hippocampal atrophy in Alzheimer's disease: A longitudinal MRI study. Brain. 1996; 119:2001–2007. [PubMed: 9010004]
26. Gee J, Reivich M, Bajcsy R. Elastically deforming a three-dimensional atlas to match anatomical brain images. J Comput Assist Tomogr. 1993; 17(2):225–236. [PubMed: 8454749]
27. Gering DT, Nabavi A, Kikinis R, Hata N, ODonnell LJ, Grimson WEL, Jolesz FA, Black PM, Wells WM III. An integrated visualization system for surgical planning and guidance using image fusion and an open MR. J Magn Reson Imag. 2001; 13:967–975.
28. Goldstein JM, Goodman JM, Seidman LJ, Kennedy DN, Makris N, Lee H, Tourville J, Caviness VS Jr, Faraone SV, Tsuang MT. Cortical abnormalities in schizophrenia identified by structural magnetic resonance imaging. Arch Gen Psychiatry. 1999; 56:537–547. [PubMed: 10359468]
29. Han X, Fischl B. Atlas renormalization for improved brain MR image segmentation across scanner platforms. IEEE Trans Med Imag. Apr; 2007 26(4):479–486.
30. Heckemann RA, Hajnal JV, Aljabar P, Rueckert D, Hammers A. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. Neuroimage. 2006; 33(1):115–126. [PubMed: 16860573]
31. Iosifescu DV, Shenton ME, Warfield SK, Kikinis R, Dengler J, Joelsz FA, McCarley RW. An automated registration algorithm for measuring MRI subcortical brain structures. Neuroimage. 1997; 6(1):13–25. [PubMed: 9245652]
32. Isgum I, Staring M, Rutten A, Prokop M, Viergever MA, van Ginneken B. Multi-atlas-based segmentation with local decision fusion-application to cardiac and aortic segmentation in CT scans. IEEE Trans Med Imag. Jul; 2009 28(7):1000–1010.
33. Jaakkola, T. Tutorial on variational approximation methods. In: Opper, M.; Saad, D., editors. Advanced Mean Field Methods: Theory and Practice. Cambridge, MA: MIT Press; 2000.
34. Jack CR, Petersen RC, Xu YC, O'Brien PC, Smith GE, Ivnik RJ, Boeve BF, Waring SC, Tangalos EG, Kokmen E. Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. Neurology. 1999; 52
35. Johnson K, Jones K, Holman B, Becker J, Spiers P, Satlin A, Albert M. Preclinical prediction of alzheimers disease using spect. Neurology. 1998; 50:1563–1571. [PubMed: 9633695]
36. Joshi S, Davis B, Jomier M, Gerig G. Neuroimage. 2004; 23:151–160.
37. Kennedy DN, Filipek PA, Caviness VS. Anatomic segmentation and volumetric calculations in nuclear magnetic resonance imaging. IEEE Trans Med Imag. Jan; 1989 8(1):1–7.
38. Killiany RJ, Gomez-Isla T, Moss M, Kikinis R, Sandor T, Jolesz F, Tanzi R, Jones K, Hyman BT, Albert MS. The use of structural MRI to predict who will get alzheimers disease. Ann Neurol. 2000; 47:430–439. [PubMed: 10762153]
39. Klein A, Andersson J, Ardekani BA, Ashburner J, Avants B, Chiang MC, Christensen GE, Collins DL, Hellier P, Song JH, Jenkinson M, Lepage C, Rueckert D, Thompson P, Vercauteren T, Woods RP, Mann JJ, Parsey RV. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. NeuroImage. 2009; 46(3)
40. Van Leemput K. Encoding probabilistic brain atlases using Bayesian inference. IEEE Trans Med Imag. Jun; 2009 28(6):822–837.
41. Van Leemput K, Maes F, Vandermeulen D, Suetens P. Automated model-based tissue classification of MR images of the brain. IEEE Trans Med Imag. Oct; 1999 18(10):897–908.
42. Lotjonen JMP, Wolz R, Koikkalainen JR, Thurfjell L, Waldemar G, Soininen H, Rueckert D. Fast and robust multi-atlas segmentation of brain magnetic resonance images. NeuroImage. 2010; 49(3):2352–2365. [PubMed: 19857578]
43. Miller MI, Christensen GE, Amit Y, Grenander U. Mathematical textbook of deformable neuroanatomies. Proc Nat Acad Sci. 1993; 90(24):11944–11948. [PubMed: 8265653]
44. Morris JC. The clinical dementia rating (CDR): Current version and scoring rules. Neurology. 1993; 43:2412–2414. [PubMed: 8232972]

45. Neal, RM.; Hinton, GE. Learning in Graphical Models. Vol. 89. Cambridge, MA: MIT Press; 1998. A view of the EM algorithm that justifies incremental, sparse, and other variants; p. 355-368.
46. Nielsen M, Florack L, Deriche R. Regularization, scale-space, and edge detection filters. J Math Imag Vision. 1997; 7(4):291–307.
47. Pizer SM, Fletcher T, Fridman Y, Fritsch DS, Gash AG, Glotzer JM, Joshi S, Thall A, Tracton G, Yushkevich P, Chaney EL. Deformable m-reps for 3D medical image segmentation. Int J Comput Vis. 2003; 55(2/3):85–106.
48. Pohl KM, Fisher J, Grimson W, Kikinis R, Wells WM. A Bayesian model for joint segmentation and registration. Neuroimage. 2006; 31:228–239. [PubMed: 16466677]
49. Pohl, KM.; Fisher, J.; Shenton, M.; McCarley, RW.; Grimson, WEL.; Kikinis, R.; Wells, WM. Proc of MICCAI. Vol. 4191. New York: Springer; 2006. Logarithm odds maps for shape representation; p. 955-963.Lecture Notes Computer Science
50. Prastawa M, Gilmore J, Lin W, Gerig G. Automatic segmentation of MR images of the developing newborn brain. Med Image Anal. 2005; 9:457–466. [PubMed: 16019252]
51. Rohlfing T, Brandt R, Menzel R, Maurer CR. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. NeuroImage. 2004; 21(4):1428–1442. [PubMed: 15050568]
52. Rohlfing T, Maurer CR Jr. Shaped-based averaging. IEEE Trans Med Imag. Jan; 2007 16(1):153–161.
53. Rohlfing T, Russakoff DB, Mauerer CR. Performancebased classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. IEEE Trans Med Imag. Aug; 2004 23(8):983–994.
54. Sabuncu MR, Balci S, Shenton ME, Golland P. Imagedriven population analysis through mixture-modeling. IEEE Trans Med Imag. Sep; 2009 28(9):1473–1487.
55. Sabuncu, MR.; Yeo, BTT.; Van Leemput, K.; Fischl, B.; Golland, P. Supervised nonparameteric image parcellation. Proc. of MICCAI; 2009; New York: Springer; 2009. p. 1075-1083.Lecture Notes Computer Science
56. Sabuncu, MR.; Yeo, BTT.; Van Leemput, K.; Vercauteren, T.; Golland, P. Proc of MICCAI. Vol. 5761. New York: Springer; 2009. Asymmetric image template registration; p. 565-573.Lecture Notes Computer Science
57. Sabuncu, MR.; Yeo, BTT.; Van Leemput, K.; Fischl, B.; Golland, P. Non-parameteric mixture models for supervised image parcellation. Workshop Probabilistic Models Medical Image Analysis, Proc. Int. Conf. Med. Image Computing Computer Assist. Intervent (MICCAI); 2009.
58. Sdika M. Combining atlas based segmentation and intensity classification with nearest neighbor transform and accuracy weighted vote. Med Image Anal. 2010; 14(2):219–226. [PubMed: 20056473]
59. Seidman LJ, Faraone SV, Goldstein JM, Goodman JM, Kremen WS, Toomey R, Tourville J, Kennedy D, Makris N, Caviness VS. Thalamic and amygdala-hippocampal volume reductions in first-degree relatives of patients with schizophrenia: An MRI-based morphometric analysis. Biol Psychiatry. 1999; 46:941–954. [PubMed: 10509177]
60. Spiridon M, Fischl B, Kanwisher N. Location and spatial profile of category-specific regions in human extrastriate cortex. Human Brain Mapp. 2005; 27(1):77–89.
61. Thirion J-P. Image matching as a diffusion process: An analogy with Maxwells demons. Med Image Anal. 1998; 2(3):243–260. [PubMed: 9873902]
62. Tu Z, Narr KL, Dollár P, Dinov I, Thompson PM, Toga AW. Brain anatomical structure segmentation by hybrid discriminative/generative models. IEEE Trans Med Imag. Apr; 2008 27(4):495–508.
63. Twining, CJ.; Cootes, T.; Marsland, S.; Petrovic, V.; Schestowitz, R.; Taylor, CJ. Proc IPMI. Vol. 3565. New York: Springer; 2005. A unified informationtheoretic approach to groupwise non-rigid registration and model building; p. 1-14.Lecture Notes Computer Science
64. Van Leemput K, et al. Automated model-based bias field correction of MR images of the brain. IEEE Trans Med Imag. Oct; 1999 18(10):885–896.

65. Vercauteren, T., et al. Proc of MICCAI. Vol. 5241. New York: Springer; 2008. Symmetric log-domain diffeomorphic registration: A demons-based approach; p. 754-761.Lecture Notes Computer Science

66. Wang Q, Chen L, Yap P-T, Wu G, Shen D. Groupwise registration based on hierarchical image clustering and atlas synthesis. Human Brain Mapp. 200910.1002/hbm.20923

67. Warfield S, et al. Simultaneous truth and performance level estimation (STAPLE): An algorithm for validation of image segmentation. IEEE Trans Med Imag. Jul; 2004 23(7):903–921.

68. Wells WM, Kikinis R, Grimson WEL, Jolesz F. Adaptive segmentation of MRI data. IEEE Trans Med Imag. 1996; 15:429–442.

69. Freesurfer Wiki [Online]. Available: http://surfer.nmr.mgh.harvard.edu

70. Yang J, Staib LH, Duncan JS. Neighbor-constrained segmentation with level set based 3D deformable models. IEEE Trans Med Imag. Aug; 2004 23(8):940–948.

71. Yeo BTT, Sabuncu M, Vercauteren T, Ayache N, Fischl B, Golland P. Spherical demons: Fast diffeomorphic landmark-free surface registration. IEEE Trans Med Imag. Mar; 2010 29(3):650–668.

72. Yeo BTT, Sabuncu MR, Desikan R, Fischl B, Golland P. Effects of registration regularization and atlas sharpness on segmentation accuracy. Med Image Anal. 2008; 12(5):603–615. [PubMed: 18667352]

73. Yeo BTT, Vercauteren T, Fillard P, Pennec X, Golland P, Ayache N, Clatz O. DT-REFinD: Diffusion tensor registration with exact finite-strain differential. IEEE Trans Med Imag. Dec; 2009 28(12):1914–1928.

74. Zijdenbos A, Forghani R, Evans A. Automatic pipeline analysis of 3-D MRI data for clinical trials: Application to multiple sclerosis. IEEE Trans Med Imag. Oct; 2002 21(10):1280–1291.

Appendix A. Pairwise Registration

In deriving the segmentation algorithms in this paper, we make a crucial simplifying assumption that the warps $\{\Phi_n\}$ that map individual training images to the test image coordinates are observed. In practice, we compute these warps via a pairwise registration algorithm. Over the past decade, a wide variety of registration algorithms have been proposed in the literature. The optimal choice of a registration algorithm remains an open question that can be partially guided by a recent study that compares a broad set of pairwise registration algorithms [39]. Here, we make our choice based on the following three criteria.

1. *Speed and computational efficiency.* Since the test subject must be registered with each training subject, we need to perform registration many times. Recent algorithms based on Thirion's Demons algorithm [61] yield fast and efficient inter-subject registration.
2. *Rich deformation model.* The quality of the segmentation results depends on the accuracy of the alignment. The Klein *et al.* study [39] found that a rich, dense deformation model, in general, yields better alignment accuracy. In this work, we rely on a particular parametrization of diffeomorphic—smooth and invertible—transformations.
3. *SSD similarity measure.* To be consistent with our choice of the image likelihood model in (6), we use the sum of squared differences (SSD) as the similarity measure. As we discuss in Section VI, this particular choice is probably not optimal and can be improved by employing more sophisticated similarity measures, e.g., mutual information.

Based on these criteria, we choose the asymmetric bidirectional image-template registration algorithm, details of which are presented in [56]. This method is based on an efficient Demons-style algorithm [65] that uses a one-parameter subgroup of diffeomorphism. The

warp Φ is parameterized with a smooth, stationary velocity field v: $\mathbb{R}^3 \mapsto \mathbb{R}^3$ via an Ordinary Differential Equation (ODE) [4]

$$\frac{\partial \Phi(x,t)}{\partial t} = v(\Phi(x,t))$$

and the initial condition $\Phi(x, 0) = x$. The warp $\Phi(x) = \exp(v)(x)$ can be computed efficiently using scaling and squaring and inverted by using the negative of the velocity field: $\Phi^{-1} = \exp(-v)$ [4].

We impose an elastic-like regularization on the stationary velocity field

$$p(\Phi=\exp(v)) = \frac{1}{Z_\lambda} \exp\left[-\lambda \sum_{y\in\Omega} \sum_{j,k=1,2,3} \left(\frac{\partial^2}{\partial x_j^2} v_k(x)\Big|_{x=y} \right)^2 \right] \quad (24)$$

where $\lambda > 0$ is the warp stiffness parameter, Z_λ is a partition function that depends only on λ, and x_j and v_k denote the jth and kth component (dimension) of position x and velocity v, respectively. Higher values of the warp stiffness parameter λ yield more rigid warps.

To construct the registration objective function, we assume a simple additive Gaussian noise model that leads to the following optimization problem for registering the nth training image to the test subject

$$\widehat{v}^n = \underset{v}{\operatorname{argmin}} \sum_{y\in\Omega} \left[\left(I(y) - \tilde{I}_n(\exp(v)(y)) \right)^2 + 2\lambda\sigma^2 \sum_{j,k=1,2,3} \left(\frac{\partial^2}{\partial x_j^2} v_k(x)\Big|_{x=y} \right)^2 \right] \quad (25)$$

where σ^2 is the stationary noise variance, and $\Phi_n \triangleq \exp(\hat{v}^n)$.

Note that (25) warps the training image (i.e., template) which makes the model truly probabilistic (for a discussion see [2], [40]). Bi-directional or symmetric approaches that apply warps to both images seem to yield more accurate alignment [7], [9], [14], [65]. Recently, we proposed an objective function that reconciles the practical advantages of symmetric registration with the asymmetric nature of image-template registration [56]

$$\widehat{v}^n = \underset{v}{\operatorname{argmin}} \sum_{y\in\Omega} \left[I(y) - \tilde{I}_n(\exp(v)(y)) \right]^2 + \left[I(\exp(-v)(y)) - \tilde{I}_n(y) \right]^2 \det(\nabla \exp(-v)(y)) + 4\lambda\sigma^2 \sum_{j,k=1,2,3} \left(\frac{\partial^2}{\partial x_j^2} v_k(x)\Big|_{x=y} \right)^2 \quad (26)$$

where $\det(\nabla \exp(-v)(y))$ denotes the determinant of the Jacobian of the inverse transformation $\exp(-v)(\cdot)$ with respect to the spatial coordinates and arises from the change of coordinates $y = \Phi(x)$ and discretization of the continuous similarity measure $\int_{\mathbb{R}^3} [I(x) - \tilde{I}_n(\Phi(x))]^2 \, dx$. The Jacobian term is larger than 1 in regions where an area in image I is mapped to a smaller area in $\tilde{I}_n$ and less than 1 elsewhere.

We solve (26) using the log-domain Demons framework [65], which decouples the optimization of the first and second terms by introducing an auxiliary transformation. The update warp is first computed using the Gauss–Newton method. The regularization is achieved by smoothing the updated warp field. It can be shown that the smoothing kernel

corresponding to (24) can be approximated with a Gaussian; $K(x) \propto \exp(-\alpha \sum_{n=1,2,3} x_i^2)$, where $\alpha = \gamma/8\lambda\sigma^2$ and $\gamma > 0$ controls the size of the Gauss-Newton step [12], [46].

Appendix B. Derivation of the EM Algorithm

Here, we present the derivation of the EM algorithm for the Global Weighted Fusion method presented in Section IV-C. We first rewrite (1) using (4)

$$\begin{aligned}\widehat{L} &= \arg\max_L \log \sum_M p(M)p(L, I|M;\{L_n, I_n\}) \\ &= \arg\max_L \log p(L, I;\{L_n, I_n\}).\end{aligned} \quad (27)$$

Let q be any distribution on M. The objective function of (27) is bounded from below by

$$\log p(L, I;\{L_n, I_n\}) - \mathrm{KL}(q(M) \,||\, p(M|L, I;\{L_n, I_n\})) \quad (28)$$

where $\mathrm{KL}(\cdot||\cdot)$ denotes the nonnegative Kullback–Leibler (KL) divergence defined as

$$\mathrm{KL}(q(M) \,||\, p(M|L, I;\{L_n, I_n\})) = \sum_M q(M) \log \frac{q(M)}{p(M|L, I;\{L_n, I_n\})}. \quad (29)$$

Equation (28) is the negative of what is usually called the free energy [33], which we will denote $\mathcal{F}(q, L)$ since it is a function of two unknown variables: the label map (segmentation) L and the distribution q. The KL divergence achieves zero only at $q^*(M) = p(M|L, I, \{L_n, I_n\})$, in which case the negative free energy $\mathcal{F}(q^*, L)$ is merely a function of L and is equal to the objective function of (27).

By simple algebraic manipulations, the free energy can be equivalently expressed as

$$\mathcal{F}(q, L) = -H(q) - \sum_M q(M) \log p(L, I, M;\{L_n, I_n\}) \quad (30)$$

where $H(q)$ denotes the entropy of q.

One interpretation of the Expectation Maximization algorithm is based on a minimization of the free energy over the two variables q and L [45]. The free energy provides a bound on the true objective function, which is optimized via coordinate-descent by alternating between the optimization over L for fixed q and vice versa. Crucially, we can see from (29) that for a fixed $\hat{L}$, $q^*(M) = p(M|\hat{L}, I, \{L_n, I_n\})$ minimizes $\mathcal{F}(\cdot, \hat{L})$. The computation of $q^*(M)$ is called the E-step. In the M-step, the algorithm minimizes $\mathcal{F}(q^*(M), L)$ over L. Since $\mathcal{F}(q^*, \hat{L})$ is equal to the objective function evaluated at $\hat{L}$, the M-step is guaranteed to improve the objective function, effectively guaranteeing convergence to a local optimum.

We can now derive the update equations for the two steps.

E-Step

For a fixed segmentation estimate $\hat{L}^{(i-1)}$, we obtain

$$\begin{aligned} q^{(i)}(M) &= p(M|\widehat{L}^{(i-1)}, I; \{L_n, I_n\}) \\ &\propto p(\widehat{L}^{(i-1)}, I|M; \{L_n I_n\}) p(M) \\ &= p_M(\widehat{L}^{(i-1)}; \{L_n\}) p_M(I; \{I_n\}) p(M) \end{aligned} \quad (31)$$

where (31) follows from the assumptions of the generative model presented in Section II. Since, in the global model we assume $\beta \to \infty$, $p(M)$ (and therefore $q^*(M)$) is zero unless $M(x)$ is equal to a constant index $n \in \{1, \ldots, N\}$ for all $x \in \Omega$. We denote these nonzero elements with m_n. Thus, the E-step can be simplified to

$$m_n^{(i)} \propto p_n(I; I_n) p_n(\widehat{L}^{(i-1)}; L_n). \quad (32)$$

M-Step

For a given $q^{(i)}$ (or, equivalently $m_n^{(i)}$), the label maps are updated using (30)

$$\begin{aligned} \widehat{L}^{(i)} &= \arg\max_L \sum_M q^{(i)}(M) \log p(L, I, M; \{L_n, I_n\}) \\ &= \arg\max_L \sum_n \sum_{x \in \Omega} m_n^{(i)} \log p_n(L(x); L_n) \end{aligned} \quad (33)$$

where (33) uses (32) and the assumptions of the generative model presented in Section II, and omits terms that are constant with respect to L. Note that in (33), each voxel can be optimized independently.

Appendix C. Derivation of the Variational EM Algorithm

Here, we derive the variational EM algorithm for the Semi-local Weighted Fusion method presented in Section IV-D. The key difference between the EM algorithm derived in the previous section and the variational EM algorithm is due to the MRF prior $p(M)$ (Section III-C) with a finite, nonzero β that allows $M(x)$ to vary spatially, while encouraging neighboring voxels to take the same value. The coupling between neighboring voxels introduced by the MRF prior makes the M-step of the algorithm computationally intractable. One approach to compute an approximate solution is to impose a structure on q[33]. This loosens the lower bound of (28), since $q^*(M) = p(M|L, I; \{L_n, I_n\})$ may not have the structure we impose on q. In other words, variational EM relaxes the optimization problem by replacing the objective function with an easier-to-optimize loose lower bound.

In our model, we use a (standard) fully factorized structure for q

$$q(M) = \prod_{x \in \Omega} q_x(M(x)). \quad (34)$$

E-Step

First, we derive the update rule for q for fixed $\hat{L}^{(i-1)}$. Substituting (3) into (28) and omitting terms not containing M, we obtain

$$\widehat{q}^{(i)} = \underset{q}{\operatorname{argmin}}\, \mathrm{KL}(q(M) \,\|\, p(M|\widehat{L}^{(i-1)}, I; \{L_n, I_n\})) \quad (35)$$

$$= \underset{q}{\operatorname{argmin}} \sum_{M} q(M) \log \frac{q(M)}{p(\widehat{L}^{(i-1)}, I|M; \{L_n, I_n\}) p(M)} \quad (36)$$

$$= \underset{q}{\operatorname{argmin}}\, \mathbb{E}_q\left(-\log p(M)\right) + \sum_{x\in\Omega} \mathbb{E}_{q_x}\left(\log q_x(M(x)) - \log p_{M(x)}(\widehat{L}^{(i-1)}(x); L_{M(x)}) p_{M(x)}(I(x); I_{M(x)})\right) \quad (37)$$

where $\mathbb{E}_q$ denotes expectation with respect to q. Using (8), we rewrite the first term of (37) as

$$\mathbb{E}_q\left(-\beta \sum_{x\in\Omega} \sum_{y\in\mathcal{N}_X} \delta(M(x), M(y))\right) + \log Z_\beta = -\beta \sum_{x\in\Omega} \sum_{y\in\mathcal{N}_X} \langle q_x, q_y \rangle + \log Z_\beta, \quad (38)$$

$$= -\beta \sum_{x\in\Omega} \mathbb{E}_{q_x}\left(\sum_{y\in\mathcal{N}_X} q_y(M(x))\right) + \log Z_\beta \quad (39)$$

where we used the linearity of expectation and $\langle q_x, q_y \rangle$ denotes the inner product between the probability vectors q_x and q_y. Inserting (39) into (37) we obtain

$$\widehat{q}^{(i)} = \underset{q}{\operatorname{argmin}} \sum_{x\in\Omega} \mathbb{E}_{q_x}\left(\beta \sum_{y\in\mathcal{N}_X} q_y(M(x)) + \log \frac{q_x(M(x))}{p_{M(x)}(\widehat{L}^{(i-1)}(x); L_{M(x)}) p_{M(x)}(I(x); I_{M(x)})}\right). \quad (40)$$

To solve (40), we can differentiate the objective function with respect to each q_X and equate to zero. With the probability constraint, i.e., $\sum_n q_x^{(i)}(M(x)=n)=1$, $\forall X \in \Omega$, this implies (19).

M-Step

With the structure of (34), the segmentation update can be computed using (30)

$$\widehat{L}^{(i)} = \underset{L}{\operatorname{argmax}} \sum_{x\in\Omega} \sum_{n=1}^{N} q_x^{(i)}(M(x)=n) \times \log p(L(x), I(x)|M(x)=n; L_n, I_n) \quad (41)$$

$$= \underset{L}{\operatorname{argmax}} \sum_{x\in\Omega} \sum_{n=1}^{N} q_x^{(i)}(M(x)=n) \log p_n(L(x); L_n). \quad (42)$$

Since each voxel can be considered independently, we obtain (20).

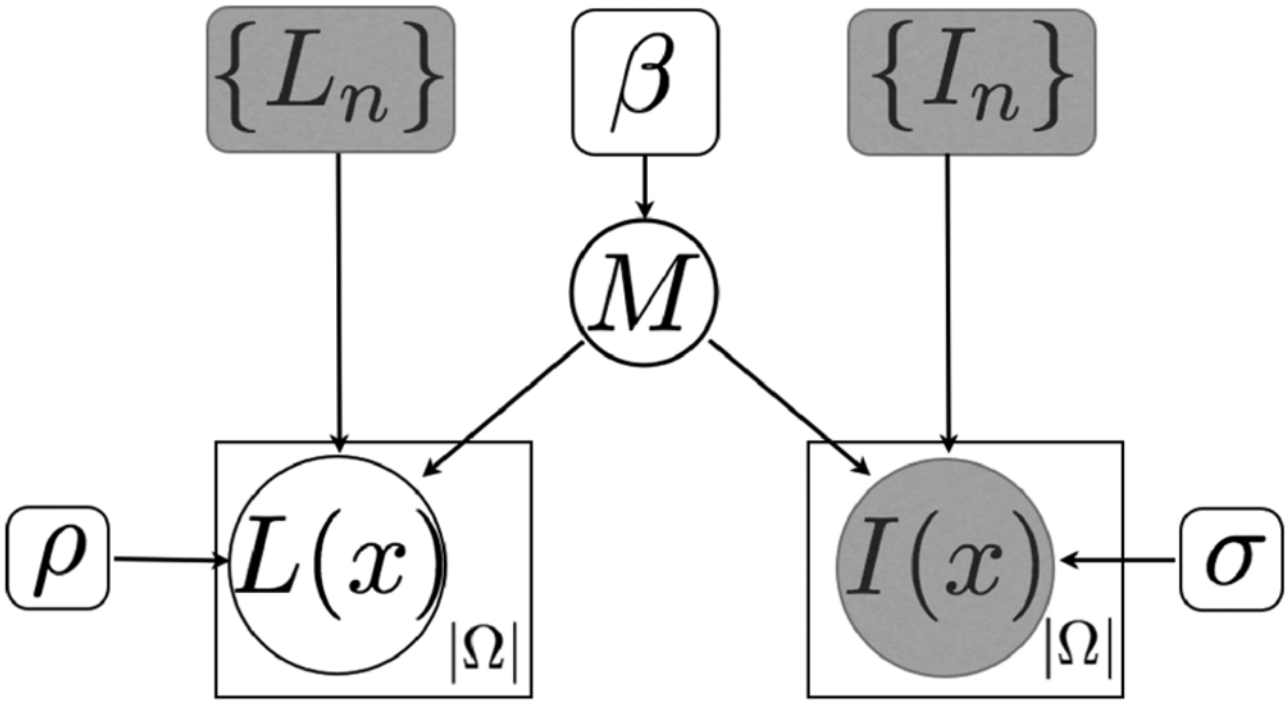

Fig. 1.

Graphical model that depicts the relationship between the variables. Squares indicate nonrandom parameters, circles indicate random variables. Replications are illustrated with plates (bounding $L(x)$ and $I(x)$). The $|\Omega|$ in the corner of the plate indicates the variables inside are replicated that many times (i.e., once for each voxel), and thus are conditionally independent. Shaded variables are observed.

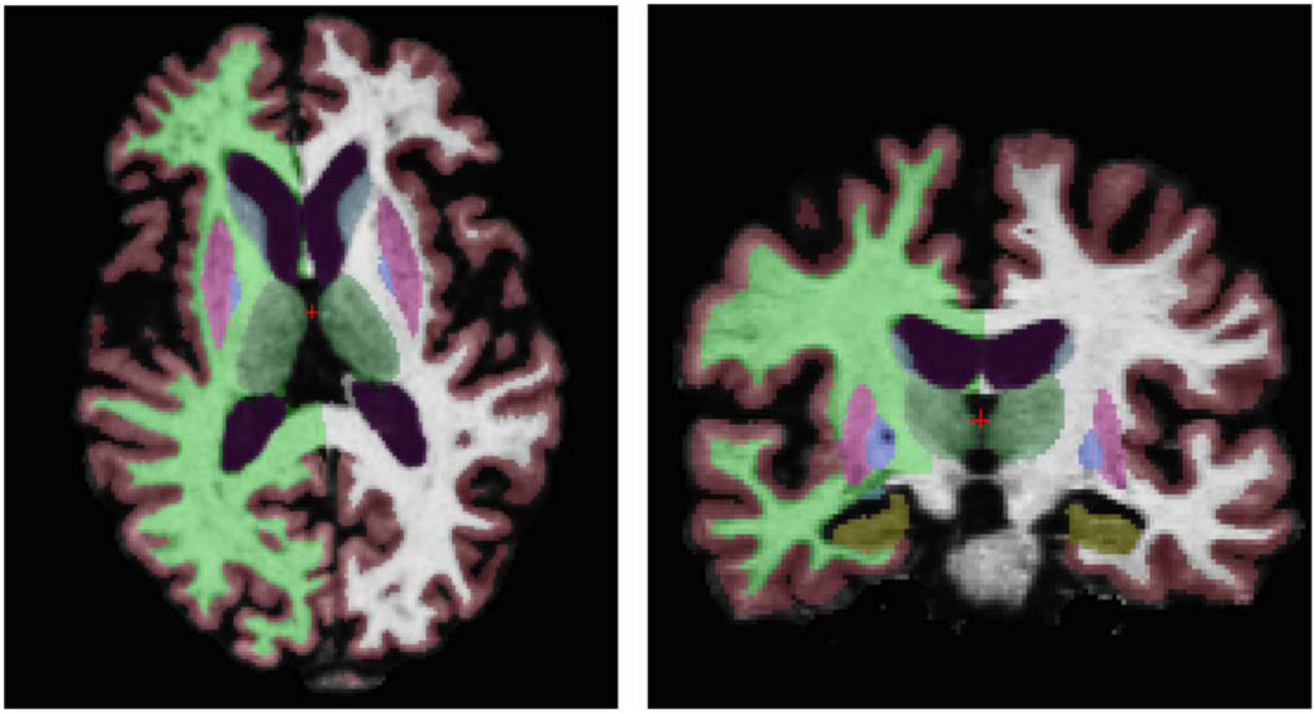

Fig. 2.
A typical segmentation obtained with the local mixture model. 2D slices are shown for visualization only. All computations are done in 3D.

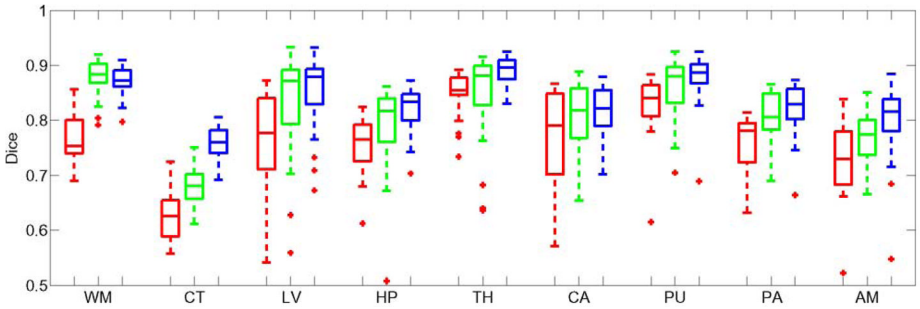

Fig. 3.

Dice scores obtained using Majority Voting and various label prior models: Nearest Neighbor (red), Tri-linear (green), and LogOdds (blue). Dice scores for the two hemispheres were averaged. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles. The whiskers extend to 2.7 standard deviations around the mean, and outliers are marked individually as a "*."

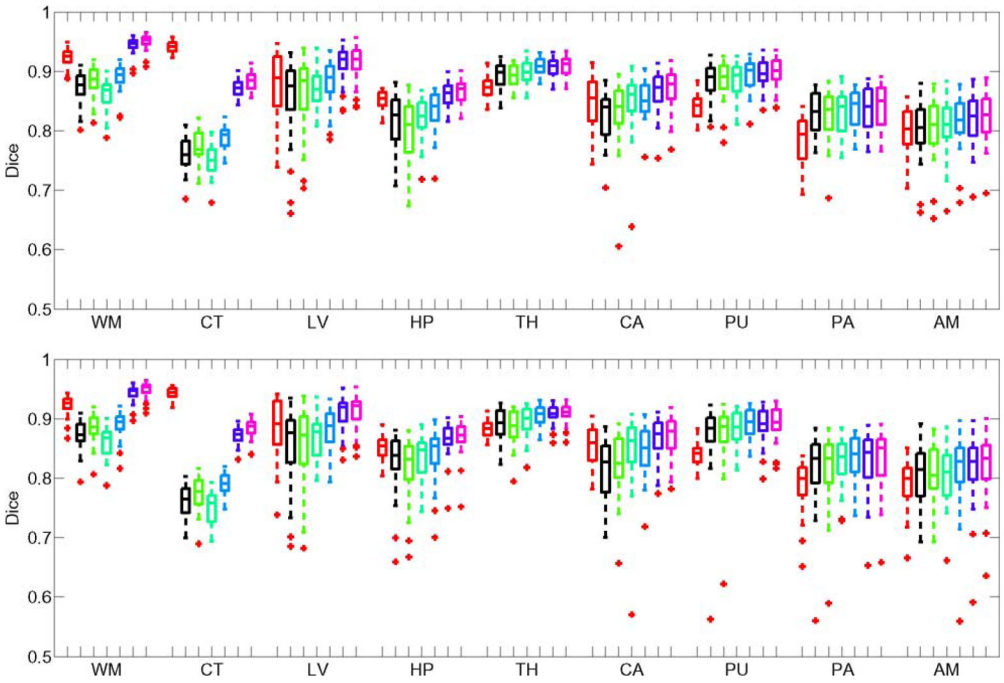

Fig. 4.

Dice scores for all methods (top: left hemisphere, bottom: right hemisphere): FreeSurfer (red), Majority Voting (black), STAPLE (light green), Majority10 (dark green), Global Weighted Fusion (light blue), Local Weighted Voting (dark blue), Semi-local Weighted Fusion (purple). On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles. The whiskers extend to 2.7 standard deviations around the mean, and outliers are marked individually as a "*."

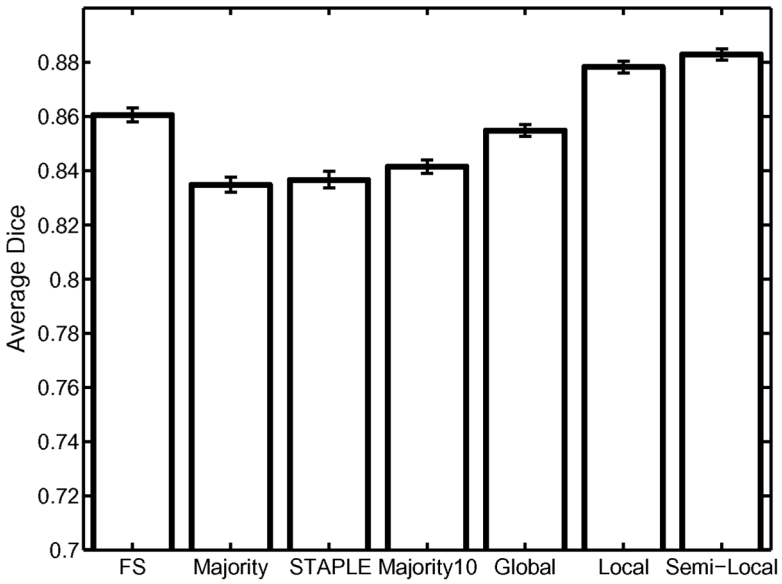

Fig. 5.

Average Dice scores for each algorithm (FS: FreeSurfer, Majority: Majority Voting, STAPLE, Majority10, Global: Global Weighted Fusion, Local: Local Weighted Voting, and Semi-Local: Semi-local Weighted Fusion). Error bars show standard error. Each subject and ROI was treated as an independent sample with equal weight.

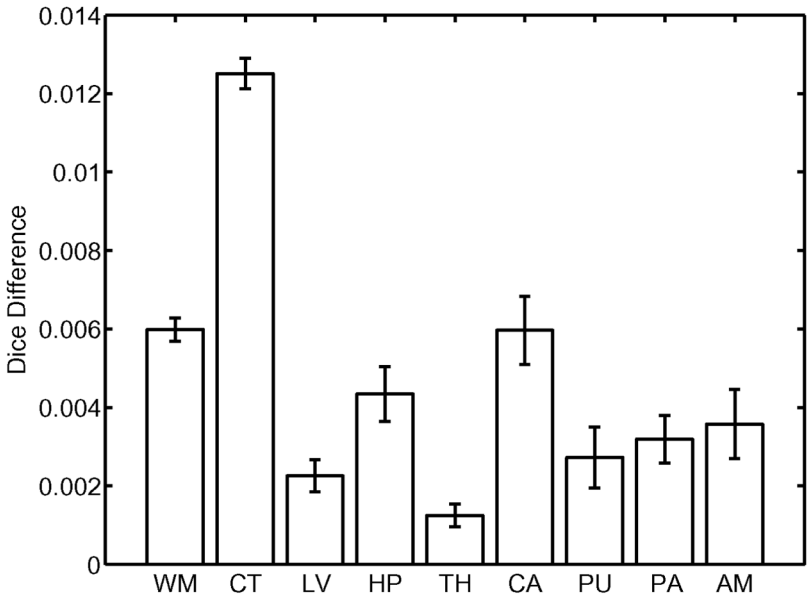

Fig. 6.

Average Dice differences: Semi-Local Weighted Fusion minus Local Weighted Voting. Overall, Semi-Local Weighted Fusion achieves better segmentation. Error bars show standard error.

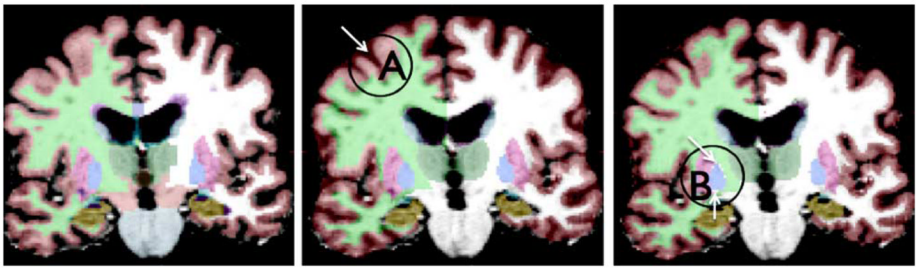

Fig. 7.

The segmentations of the subject that Semi-local Weighted Fusion performed the worst on. Left to right: FreeSurfer, Global and Semi-local Weighted Fusion. Common mistakes (indicated by arrows): (A) Global Weighted Fusion tends to over-segment complex shapes like the cortex. (B) Semi-local Weighted Fusion does not encode topological information, as FreeSurfer does. Hence it may assign an “unknown” or “background” label (white) in between the pallidum (blue), putamen (pink), and white matter (green).

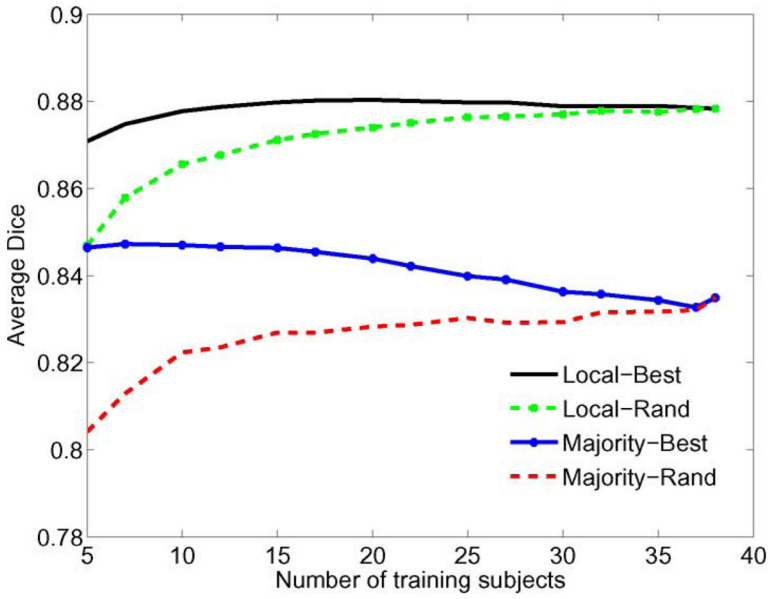

Fig. 8.

The average Dice score for Majority Voting (Majority) and Local Weighted Voting (Local) as a function of the number of training subjects. We consider two strategies to select the training subjects: (1) randomly selecting a set of training subjects (Rand), (2) selecting the best training subjects that are globally most similar to the test subject (Best). The average Dice score reaches 83.9% for Majority Voting and 87.8% for Local Weighted Voting, when all 38 subjects are used.

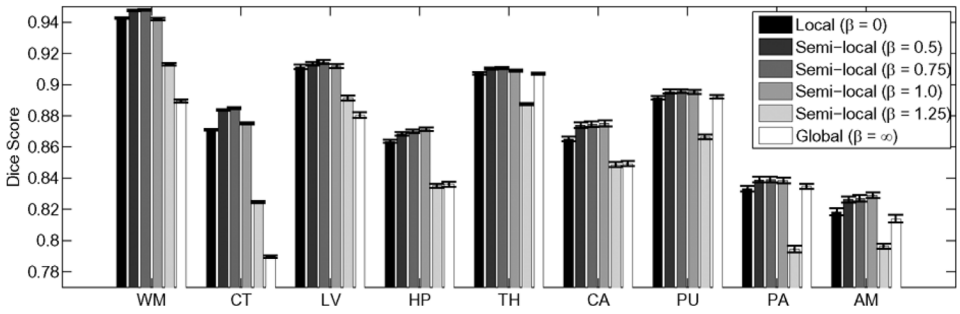

Fig. 9.

Average Dice scores for different β values in the MRF membership prior of (8). Error bars show standard error.

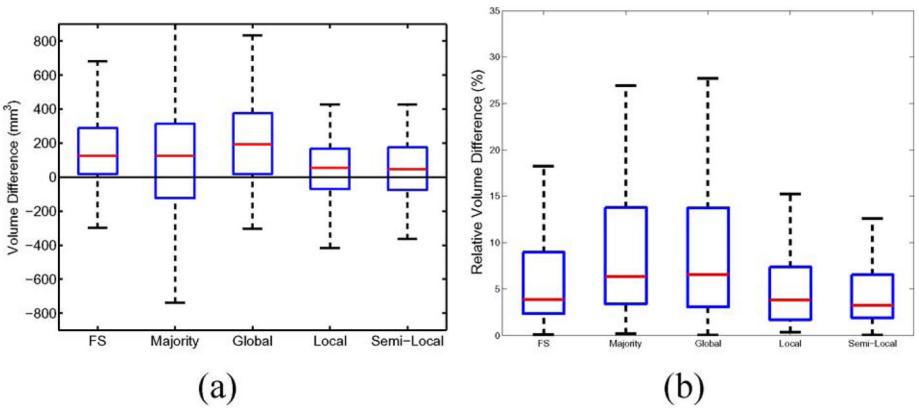

Fig. 10.

Hippocampal volume differences on the data from Experiment 1. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles. The whiskers extend to 2.7 standard deviations around the mean. (a) Automatic minus Manual volumes. (b) Relative volume differences [(23)].

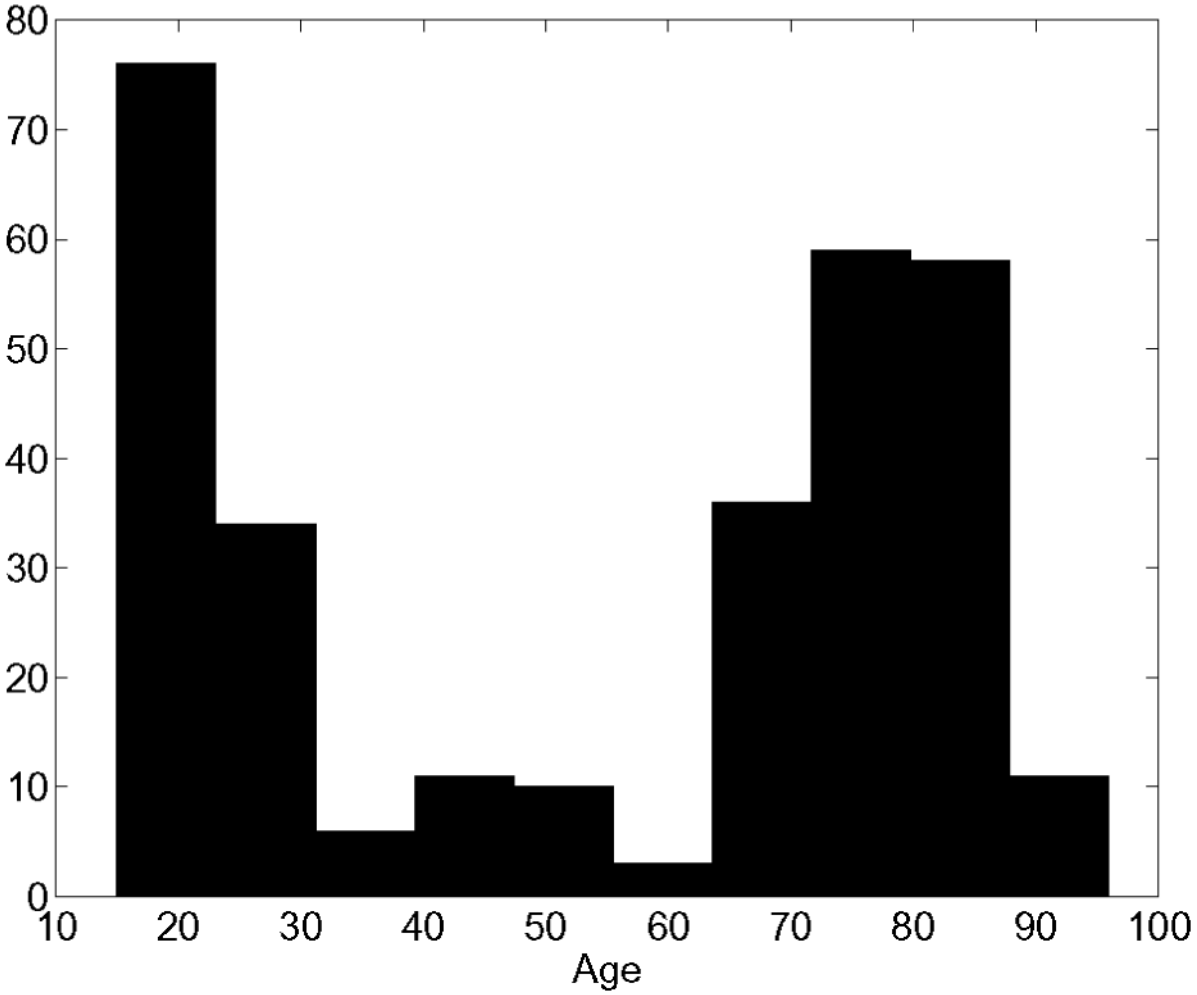

Fig. 11.
Age histogram of 282 subjects in Experiment 2.

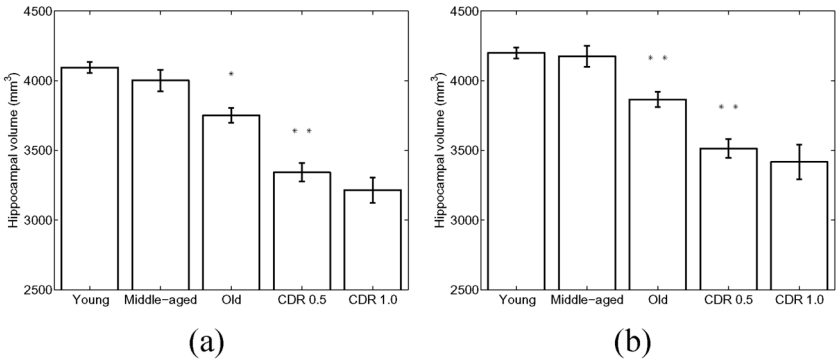

Fig. 12.

Hippocampal volumes for five different groups in Experiment 2. Error bars indicate standard error across subjects. Stars indicate that the volume measurements in the present group are statistically significantly smaller than the measurements in the neighboring group to the left. (Unpaired, single-sided t-test. * $p < 0.05$, ** $p < 0.01$). (a) Left hippocampus. (b) Right hippocampus.

TABLE I

Comparison of Average Dice Scores. **Boldface** Font Indicates Best Scores for Each ROI That Have Achieved Statistical Significance (Paired Permutation Test: $p < 0.05$, Bonferroni Corrected). *ITALIC FONT* Indicates Best Scores That are Statistically Equivalent. As a Reference, Last Row Lists Approximate Average Volumes

	WM	**CT**	**LV**	**HP**	**TH**	**CA**	**PU**	**PA**	**AM**
FreeSurfer	0.923	**0.941**	0.878	0.851	0.879	0.849	0.840	0.786	0.796
Majority Vote	0.873	0.760	0.853	0.821	0.891	0.819	0.876	0.822	0.799
STAPLE	0.884	0.774	0.857	0.807	0.889	0.827	0.881	0.821	0.792
Majority 10	0.861	0.750	0.869	0.827	0.897	0.848	0.886	0.829	0.806
Global Weighted Fusion	0.890	0.790	0.880	0.836	*0.907*	0.849	*0.892*	*0.835*	0.814
Local Weighted Fusion	*0.943*	0.871	*0.912*	0.864	*0.907*	0.865	0.891	0.833	0.818
Semi-local Weighted Fusion	*0.949*	0.884	*0.914*	**0.868**	*0.908*	**0.871**	*0.894*	*0.836*	**0.822**
Volumes ($\times 10^3 mm^3$)	450	448	25	7	14	7	10	3	3

TABLE II

Average Run-Time (in CPU Hours on a Modern Machine -Intel Xeon 3 GHz With a 32 GB RAM-) to Segment One Test Subject. The Time Taken up by Registration is Listed in Parentheses. Global: Global Weighted Fusion, Local: Local Weighted Voting, Semi-Local: Semi-Local Weighted Fusion

FreeSurfer	Majority	STAPLE	Majority 10	Global	Local	Semi-local
10(9)	24 (23)	28 (23)	8(7)	32 (23)	24 (23)	40 (23)