HHS Public Access

Author manuscript

IEEE Trans Pattern Anal Mach Intell. Author manuscript; available in PMC 2013 December 22.

Published in final edited form as:

IEEE Trans Pattern Anal Mach Intell. 2013 March ; 35(3): 611–623. doi:10.1109/TPAMI.2012.143.

Multi-Atlas Segmentation with Joint Label Fusion

Hongzhi Wang[1] **[Member, IEEE]**, **Jung W. Suh**[2], **Sandhitsu R. Das**[1], **John Pluta**[3], **Caryne Craige**[4], and **Paul A. Yushkevich**[1] **[Member, IEEE]**

[1]Penn Image Computing and Science Lab, Department of Radiology, University of Pennsylvania

[2]HeartFlow, Inc

[3]Departments of Neurology and Radiology, University of Pennsylvania

[4]School of Medicine, Temple University

Abstract

Multi-atlas segmentation is an effective approach for automatically labeling objects of interest in biomedical images. In this approach, multiple expert-segmented example images, called *atlases*, are registered to a target image, and deformed atlas segmentations are combined using *label fusion*. Among the proposed label fusion strategies, weighted voting with spatially varying weight distributions derived from atlas-target intensity similarity have been particularly successful. However, one limitation of these strategies is that the weights are computed independently for each atlas, without taking into account the fact that different atlases may produce similar label errors. To address this limitation, we propose a new solution for the label fusion problem, in which weighted voting is formulated in terms of minimizing the total expectation of labeling error, and in which pairwise dependency between atlases is explicitly modeled as the joint probability of two atlases making a segmentation error at a voxel. This probability is approximated using intensity similarity between a pair of atlases and the target image in the neighborhood of each voxel. We validate our method in two medical image segmentation problems: hippocampus segmentation and hippocampus subfield segmentation in magnetic resonance (MR) images. For both problems, we show consistent and significant improvement over label fusion strategies that assign atlas weights independently.

Index Terms

multi-atlas label fusion segmentation; dependence; hippocampal segmentation

I. Introduction

Atlas-based segmentation is motivated by the observation that segmentation strongly correlates with image appearance. A target image can be segmented by referring to atlases, i.e. expert-labeled sample images. After warping the atlas to the target image via deformable registration, one can directly transfer labels from the atlas to the target image. As an extension, multi-atlas based segmentation makes use of more than one atlas to compensate

for potential bias associated with using a single atlas and applies label fusion to produce the final segmentation. This method requires higher computational costs but, as extensive empirical studies have verified in the recent literature, e.g. [16], [3], [22], it is more accurate than single atlas based segmentation. Enabled by availability and low cost of multi-core processors, multi-atlas label fusion (MALF) is becoming more accessible to the medical image analysis community. Recently, the concept has also been applied in computer vision for segmenting natural images [37], [21].

Errors produced by atlas-based segmentation can be attributed to dissimilarity in the structure (e.g., anatomy) and appearance between the atlas and the target image. Recent research has been focusing on addressing this problem. For instance, such errors can be reduced by optimally constructing a single atlas that is the most representative of the population using training data [12], [11], [18]. Constructing multiple representative atlases from training data has been considered as well and usually works better than single-atlas based approaches. Multi-atlas construction is done either by constructing one representative atlas for each mode obtained from clustering training images [5], [2], [32] or by simply selecting the most relevant atlases for the unknown image on-the-fly [30], [1], [41]. Either way, one needs to combine the segmentation results obtained by referring to different atlases to produce the final solution.

Most existing label fusion methods are based on weighted voting, [30], [16], [3], [17], [33], where each atlas contributes to the final solution according to a non-negative weight, with atlases more similar to the target image receiving larger weights. Among weighted voting methods, those that derive weights from local similarity between the atlas and target, and thus allow the weights to vary spatially, have been most successful in practice [3], [17], [33]. One common property of these spatially variable weighted voting MALF methods is that the weights for each atlas are computed independently, only taking into consideration the similarity between the warped atlas and the target image. As such, these methods are less effective when the label errors produced by the atlases are not independent, e.g. most atlases produce similar errors. As a simple example, suppose that a single atlas is duplicated multiple times in the atlas set. If weights are derived only from atlas-target similarity, the total contribution of the repeated atlas to the consensus segmentation will increase in proportion to the number of times the atlas is repeated, making it more difficult to correct the label error produced by the duplicated atlas. Likewise, if the atlas set is dominated by a certain kind of anatomical feature or configuration, there will be an inherent bias towards that feature, even when segmenting target images which do not share that feature. As the result, the quality of the segmentation for the less frequent anatomical features/ configurations may be reduced.

Another class of label fusion methods perform majority voting among a small subset of atlases that globally or locally best match the target image, discarding the information from poor matching atlases [3], [7]. These methods are less susceptible to the problem described, since the atlas appearing multiple times would only be included in the voting if it is similar to the target image. However, by completely discarding information from poorer matches, these methods lose the attractive property of voting arising from the central limit theorem. In particular, when all atlases are roughly equally similar to the target image, performing

voting only among the few best atlases will have greater expected error than voting between all atlases.

This paper derives a novel label fusion strategy that aims to reduce the bias due to the fact that atlases may produce correlated segmentation errors, without sacrificing the attractive properties of voting. The strategy is derived from formulating the weighted voting problem as an optimization problem over unknown voting weights, with the expected total error of the consensus segmentation relative to the unknown true segmentation being minimized. This formulation requires the joint distribution of label errors produced by any pair of atlases in the neighborhood of each voxel to be known. In practice, this distribution is unknown, and we estimate it using image intensity similarity. However, unlike previous methods, similarity with the target image is not measured independently at each atlas. Instead, intensity similarity between the target and each pair of images is considered, which leads to an ability to explicitly estimate the probability that a pair of atlases produce the same segmentation error. We hypothesize that this strategy will improve segmentation accuracy over existing techniques that consider atlas-target similarity independently [3], [33]. To test this hypothesis, we perform cross-validation segmentation experiments in manually labeled MRI datasets, and report significant improvements over earlier methods.

Preliminary versions of this work appeared in [39], [38].

II. Multi-Atlas Based Segmentation

We start with a brief overview of MALF. Let F_T be a target image to be segmented and $A_1 = (F_1, S_1),\ldots,A_n = (F_n, S_n)$ be n atlases. F_i and S_i denote the i_{th} warped atlas image and the corresponding warped manual segmentation of this atlas, obtained by performing deformable image registration to the target image. Each atlas produces one candidate segmentation for the target image. Each of the candidate segmentations may contain some segmentation errors. Label fusion is the process of integrating the candidate segmentations produced by all atlases to improve the segmentation accuracy in the final solution.

Errors produced in atlas-based segmentation are mainly due to registration errors, i.e. registration associates wrong regions from an atlas to the target image. Under the assumption that the errors produced by using different atlases are not identical, employing multiple atlases can effectively reduce label errors. For example, the majority voting method [13], [19] simply counts the votes for each label from each warped atlas and chooses the label receiving the most votes to produce the final segmentation $\hat{S}_T$:

$$\hat{S}_T(x) = \operatorname{argmax}_{l \in \{1 \ldots L\}} \sum_{i=1}^{n} S_i^l(x) \quad (1)$$

where l indexes through labels and L is the number of all possible labels, x indexes through image pixels. $S_i^l(x)$ is the vote for label l produced by the i_{th} atlas, defined by:

$$S_i^l(x) = \begin{cases} 1 & \text{if } S_i(x) = l; \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

The power of voting in removing independent noise has been long recognized. For instance, for a simple binary (yes/no) voting problem, in a group of 23 voters, if three voters always give the same vote and the remaining voters vote randomly, the chance that the final voting result is consistent with what the three resolute voters choose is ~75% [28]. For multi-way voting problems, majority voting is even more powerful in removing independent noise. In our problem, suppose that atlas A_i produces correct labels for the target image with probability p_i. The probability that the atlas will produce any particular wrong label can be roughly estimated by $(1-p_i)/(L-1)$. When $p_i > (1-p_i)/(L-1)$, the atlas works better than random guess. When segmentation errors produced by different atlases are independent, the probability that multiple atlases agree on the same wrong label is exponentially suppressed compared to the probability that they agree on the same correct label. Hence, the combined results are expected to produce significantly fewer errors than those produced by any single atlas.

Since majority voting assigns equal weights to different atlases, it makes a strong assumption that different atlases produce equally accurate segmentations for the target image. However, as a complex optimization problem, the performance of deformable registration is sensitive to the input images. Hence, it is common that different atlases may produce different registration qualities, therefore segmentations with different qualities, for the same target image.

To improve label fusion accuracy, recent work focuses on developing segmentation quality estimations based on local appearance similarity and assigning greater weights to more accurate segmentations. For instance, the votes received by label l can be estimated by:

$$\hat{S}_T^l(x)=\sum_{i=1}^{n} w_i(x) S_i^l(x) \quad (3)$$

$w_i(x)$ is a local weight assigned to the i_{th} atlas, with $\sum_{i=1}^{n} w_i(x)=1$. One way to estimate the weight is based on local image similarity under the assumption that images with similar appearance are more likely to have similar segmentations. When the summed squared distance (SSD) and a Gaussian weighting model are used [33][1], the weights can be estimated by:

$$w_i(x)=\frac{1}{Z(x)} e^{-\Sigma_{y\in\mathcal{N}(x)}[F_T(y)-F_i(y)]^2/\sigma} \quad (4)$$

where $\mathcal{N}(x)$ defines a neighborhood around x and $Z(x)$ is a normalization constant. In our experiment, we use a $(2r+1)\times(2r+1)\times(2r+1)$ cube-shaped neighborhood specified by the radius r. Since segmentation quality usually is nonuniform over the entire image, the estimation is applied based on local appearance dissimilarity. The inverse distance weighting has been applied as well [3], [17][2]:

[1] [33] proposes a general method with multiple specific implementations. Here, we refer to the "local weighted voting" implementation, i.e. equation (6) and (10) in [33].

[2] The equation corresponds to the local weighted voting with mean squared distance metric (LWV-MSD) in [3]. This paper also evaluated other metrics and found that MSD was a top performer.

$$w_i(x)=\frac{1}{Z(x)}\left[\sum_{y\in\mathcal{N}(x)}(F_T(y)-F_i(y))^2\right]^{-\beta} \quad (5)$$

where σ and β are model parameters controlling the weight distribution.

Despite the highly competitive performance produced by these image similarity based local weighted voting approaches [16], [3], [22], the key limitation of these methods is that they assign voting weights to each atlas independently, and thus cannot account for the fact label errors produced by different atlases may be correlated, as pointed out in the Introduction. Next, we introduce a method to address this limitation.

III. Joint Label Fusion

For simplicity, in the theoretical exposition that follows, we consider binary segmentation, i.e. segmentation into foreground and background labels. We assume that each voxel in the target image is labeled 0 or 1, and that each atlas segmentation also assigns 0 or 1 to each voxel. Probabilistic segmentation (where each voxel is assigned a probability of having a given label) can also be achieved in practice by using the same weighting scheme as we develop below. Likewise, a segmentation problem with more than two labels can be decomposed into multiple binary segmentation problems, i.e. segmenting each label from the remaining labels. Our method can be applied to multi-label segmentation problems by producing weight maps as described below, using weighted voting to compute a consensus segmentation for each label, and selecting at each voxel the label with the highest value of the consensus segmentation.

In binary segmentation, we can model segmentation errors produced in atlas-based segmentation as follows:

$$S_T(x)=S_i(x)+\delta^i(x) \quad (6)$$

where $\delta^i(x)$ is the label difference between the i_{th} atlas and the target image at x. $\delta^i(x) \in \{-1, 0\}$ when $S_i(x) = 1$ and $\delta^i(x) \in \{0, 1\}$ when $S_i(x) = 0$. We model the label difference as a discrete random variable, characterized by the following distribution:

$$q^i(x)=p(|\delta^i(x)|=1|F_T, F_1, \ldots, F_n) \quad (7)$$

We adopt the weighted voting framework, where at each x, a consensus segmentation $\overline{S}(x)$ is generated as the weighted sum

$$\overline{S}(x)=\sum_{i=1}^{n} w_i(x)S_i(x) \quad (8)$$

where $w_i(x)$ are spatially varying weight maps that add up to 1 at each x. Note that whereas the candidate and target segmentations are taken to be binary, the consensus segmentation $\overline{S}$

(x) is not. Our aim is to find the set of voting weights that minimize the total expected error between $S(x)$ and the true segmentation $S_T(x)$, given by

$$E_{\delta^1(x),\ldots,\delta^n(x)}\left[\left(S_T(x)-\overline{S}(x)\right)^2|F_T,F_1,\ldots,F_n\right]= \quad (9)$$

$$=E_{\delta^1(x),\ldots,\delta^n(x)}\left[\left(\sum_{i=1}^{n}w_i(x)\delta^i(x)\right)^2|F_T,F_1,\ldots,F_n\right]=\sum_{i=1}^{n}\sum_{j=1}^{n}w_i(x)w_j(x)E_{\delta^i(x),\delta^j(x)}[\delta^i(x)\delta^j(x)|F_T,F_1,\ldots,F_n] \quad (10)$$
$$=\mathbf{w}_x^t M_x \mathbf{w}_x,$$

where $\mathbf{w}_x = [w_1(x);\ldots;w_n(x)]$, and t stands for transpose. M_x is a pairwise dependency matrix with:

$$M_x(i,j)=E_{\delta^i(x)\delta^j(x)}[\delta^i(x)\delta^j(x)|F_T,F_1,\ldots,F_n] \quad (11)$$

$$=p(\delta^i(x)\delta^j(x)=1|F_T,F_1,\ldots,F_n). \quad (12)$$

$M_x(i, j)$ estimates how likely atlases i and j are to both produce wrong segmentations for the target image, given the observed feature images. Note that the product $\delta^i(x)\delta^j(x)$ can only take values 0 or 1, with $\delta^i(x)\delta^j(x) = 1$ if and only if both atlases produce a label different from the target segmentation.

Under this formulation, to achieve optimal label fusion, the voting weights should be selected such that the expectation of the combined label difference is minimized, i.e.,

$$\mathbf{w}_x^*=\underset{\mathbf{w}_x}{\operatorname{argmin}}\mathbf{w}_x^t M_x \mathbf{w}_x \text{subject to}\sum_{i=1}^{n}\mathbf{w}_x(i)=1 \quad (13)$$

Using Lagrange multipliers, we can derive a closed-form solution to this minimization problem, given by

$$\mathbf{w}_x=\frac{M_x^{-1}1n}{1_n^t M_x^{-1}1_n} \quad (14)$$

where $1_n = [1; 1;\ldots; 1]$ is a vector of size n. When M_x is not full rank, the weights can be estimated using quadratic programming optimization [27]. However, the weights that minimize (13) are not unique. We take an alternative solution by always adding an identity matrix weighted by a small positive number α to M_x. With the conditioning matrix, we minimize the following objective function instead:

$$\mathbf{w}_x^t(M_x+\alpha I)\mathbf{w}_x=\mathbf{w}_x^t M_x \mathbf{w}_x+\alpha\|\mathbf{w}_x\|_2 \text{subject to}\sum_{i=1}^{n}\mathbf{w}_x(i)=1 \quad (15)$$

Hence, adding a small conditioning identity matrix can be interpreted as enforcing a regularization term that prefers more similar voting weights assigned to different atlases.

To make sure that the added conditioning matrix is sufficient to avoid inverting an ill-conditioned matrix and the resulting voting weights also give a solution close to the global minimum of the original objective function, $\mathbf{w}_x^t M_x \mathbf{w}_x$, α should be chosen with respect to the scale of the estimated dependency matrix M_x. We found that setting $\alpha \simeq$ 1–2% of the maximal scale of estimated M_x works well. In our experiments, we estimate M_x using normalized intensity patches and the estimated M_x is in the range of [0, 4] (see below). We apply conditioning identity matrices with a fixed weight α=0.1 in all of our experiments.

A. Toy example

Suppose that a pair of atlases A_1 and A_2 produce statistically independent label errors for a given target image. If A_1 produces a wrong label 50% of the time and A_2 produces a wrong label 20% of the time, we have

$$M_x = \begin{bmatrix} 0.5 & 0.1 \\ 0.1 & 0.2 \end{bmatrix}.$$

The optimal voting weights computed by (14) are $w = [0.2, 0.8]^t$. Using (14) with a conditioning matrix with $\alpha = 0.01$ produces a solution $w = [0.2115, 0.7885]^t$. By contrast, if we compute weights independently for each atlas, e.g., $w_i \propto p(|\delta_i| = 1)^{-1}$, we obtain $w = [2/7, 5/7]^t$. The expected total segmentation errors (9) for these three weight vectors are 0.18, 0.1801, and 0.1837, respectively.

Now suppose that another atlas A_3, which is identical to A_1, is added to the atlas library. A_1 and A_3 produce identical label errors for the target image, i.e., $p(\delta_1(x)\delta_3(x) = 1) = p(|\delta_1(x)| = 1) = p(|\delta_3(x)| = 1)$. Then

$$M_x = \begin{bmatrix} 0.5 & 0.1 & 0.5 \\ 0.1 & 0.2 & 0.1 \\ 0.5 & 0.1 & 0.5 \end{bmatrix},$$

and the optimal voting weights are not unique any more, but obey the following constraint: $w_1 + w_3 = 0.2$ and $w_2 = 0.8$. The total weight assigned to the duplicated atlas remains the same. Using (14) with a conditioning matrix with $\alpha = 0.01$ produces a solution $w = [0.1068, 0.7864, 0.1068]^t$. By contrast, if we compute weights independently for each atlas, e.g., $w_i \propto p(|\delta_i| = 1)^{-1}$, we obtain the weight vector $[2/9, 5/9, 2/9]^t$, with the weight for A_2 substantially reduced and the weight for the repeated atlas A_1/A_3 boosted. The total expected errors for the three weighting strategies are 0.18, 0.1801 and 0.2099, respectively. Thus, unlike the scheme that assigns weights to atlases independently, the accuracy of the proposed scheme does not suffer from adding an atlas with redundant information. The same holds true even if atlases A_1 and A_3 are not identical, but strongly correlated.

B. Estimating M_x from intensity similarity

Our approach relies on knowing M_x, the matrix of expected pairwise joint label differences between the atlases and the target image. Note that these terms are conditioned on the target

image and all atlas images. Assuming that given the target image and the atlas images in consideration, the pairwise joint label difference term is conditionally independent from the remaining atlas images, we simplify the term as follows:

$$M_x(i,j)=p(\delta^i(x)\delta^j(x)=1|F_T,F_1,\ldots,F_n)=p(\delta^i(x)\delta^j(x)=1|F_T,F_i,F_j) \quad (16)$$

Furthermore, by assuming that given the image patches centered around the location in consideration, the pairwise joint label difference term is conditionally independent from distant voxels, we have:

$$p(\delta^i(x)\delta^j(x)=1|F_T,F_i,F_j)=p(\delta^i(x)\delta^j(x)=1|\{F_T(y),F_i(y),F_j(y)|y\in\mathcal{N}(x)\}), \quad (17)$$

In the image registration and segmentation literature, it is a common practice to use local image information between two images to predict their label difference. To make our method more comparable to previous label fusion methods, we propose to adapt the inverse distance function (5) to estimate the probability of pairwise joint label difference as follows:

$$M_x(i,j)=p(\delta^i(x)\delta^j(x)=1|\{F_T(y),F_i(y),F_j(y)|y\in\mathcal{N}(x)\})\propto\left[\sum_{y\in\mathcal{N}(x)}|F_T(y)-F_i(y)||F_T(y)-F_j(y)|\right]^{\beta}. \quad (18)$$

The constant of proportionality in (18) is irrelevant, since multiplying M_x by a positive constant does not change the solution w. Note that when $i = j$,

$$M_x(i,i)=p(\delta^i(x)^2=1|\{F_T(y),F_i(y),|y\in\mathcal{N}(x)\})\propto\left[\sum_{y\in\mathcal{N}(x)}(F_T(y)-F_i(y))^2\right]^{\beta}.$$

i.e., $M_x(i, i)$ is the inverse of the voting weight defined by inverse distance weighting function (5). Intuitively, our approximation is based on the assumption that the expectation of the label difference produced by one atlas is large when the image intensity difference between the warped atlas and the target image is large. Similarly, the expectation of any two atlases both producing a label difference is large only when both atlases have large intensity differences from the target image and the error patterns are strongly correlated.

C. Refining label fusion by local patch search

Registration errors (i.e., failure by the registration algorithm to correctly recover correspondences between objects in images) are the principal source of error in MALF. Because of the regularization constraints involved in registration, and for other reasons, such as failure to reach a global optimum of the registration objective function, the correspondences computed by registration may not always give maximum local similarity between image patches. That is, given a patch $F_T[\mathcal{N}(x)]$ in the target image and the patch $F_i[\mathcal{N}(x)]$ in the i-th registered atlas image, it is often possible to find a nearby point x' such that $F_T[\mathcal{N}(x)]$ is more similar to the patch $F_i[\mathcal{N}(x')]$ than to the patch $F_i[\mathcal{N}(x)]$. As shown recently in [8], [39], MALF performance can be moderately improved by using the

displaced patch $F_i[\mathcal{N}(x')]$ for computing the consensus segmentation of the target image. This local patch search technique can be viewed as refining the point-to-point correspondences computed by registration, while relaxing the regularization constraints that registration imposes on deformation fields.

Motivated by this observation, we determine the *local search correspondence map* between the atlas *i* and the target image as follows:

$$\xi_i(x)=\arg\min_{x'\in\mathcal{N}'(x)}\|F_i(\mathcal{N}(x'))-F_T(\mathcal{N}(x))\|^2. \quad (19)$$

Note that the domain of the minimization above is restricted to a neighborhood $\mathcal{N}'(x)$. Again, we use a cubic neighborhood definition, specified by a radius r_s. Note that $\mathcal{N}'$ and $\mathcal{N}$ may represent different neighborhoods. Given the set of local search correspondence maps $\{\xi_i\}$, we refine the definition of the consensus segmentation (8) as

$$\overline{S}_T(x)=\sum_{i=1}^{n}w_i(\xi_i(x))S_i(\xi_i(x)) \quad (20)$$

To search for the most similar image patches, larger search windows $\mathcal{N}'$ are more desirable. However, using larger searching windows more severely compromises the regularization constraint on the deformation fields, which makes the task of predicting label differences from local appearance similarities more ambiguous. As a result, the approximation (18) may become less effective. It is reasonable to expect an optimal search range that balances these two factors.

IV. Experiments

In this section, we apply our method to two segmentation problems using two types of magnetic resonance (MR) images. The first problem is whole hippocampal segmentation using T1-weighted MRI. We choose this problem because hippocampus segmentation is one of the most studied problems in brain image analysis. The hippocampus plays an important role in memory function [36]. Macroscopic changes in brain anatomy, detected and quantified by magnetic resonance imaging (MRI), consistently have been shown to be predictive of AD pathology and sensitive to AD progression [34], [9]. Accordingly, automatic hippocampus segmentation from MR images has been widely studied e.g. [6], [24], [29]. On the other hand, the hippocampus is not a homogeneous structure. It contains several distinct subfields with different roles and susceptibilities to pathology. A number of recent studies (see overview of the literature in [23]), have proposed imaging techniques and manual segmentation protocols aimed at accurately measuring hippocampal subfield volumes. Clinical utility of hippocampal subfield volumetry was recently demonstrated in dementia (e.g. [25], [26]) and other brain diseases. Hence, automatic hippocampal subfield segmentation is attracting more attention. In the second experiment, we apply our label fusion method to hippocampal subfield segmentation using focal T2-weighted MRI.

Since recent empirical studies, e.g. [3], [33], have shown that image similarity based local weighted voting is the most effective label fusion approach compared with other benchmark segmentation tools such as STAPLE [40] and FreeSurfer [11], in our experiments we focus on comparing our joint label fusion method (LWJoint) with similarity-based local weighted label fusion. For local weighted label fusion, we apply Gaussian weighting (4) (LWGaussian) and inverse distance weighting (5) (LWInverse). We also use majority voting (MV) and STAPLE [40] to define the baseline performance.

Our method has three free parameters: r, the radius of the local appearance window $\mathcal{N}$ used in similarity-based M_x estimation; r_s, the radius of the local searching window $\mathcal{N}'$ used in remedying registration errors; and β, the parameter used to transfer image similarities in the pairwise joint label difference term (18). For each segmentation experiment, the parameters are optimized by exhaustive search among a range of values in each parameter ($r \in \{1, 2, 3\}$ for whole hippocampus segmentation and $r \in \{3, 4, 5\}$ for subfield segmentation; $r_s \in \{0, 1, 2, 3\}$; $\beta \in \{0.5, 1,\ldots, 10\}$) using the atlases in a leave-one-out cross-validation strategy. We measure the average overlap between the automatic segmentation of each atlas obtained via the remaining atlases and the reference segmentation of that atlas, and find the optimal parameters that maximize this average overlap.

The optimal local appearance window and optimal local searching window are determined for LWGaussian and LWInverse methods using cross-validation as well. In addition, the optimal parameters for the weight assignment models are also determined for LWGaussian and LWInverse, with the searching range $\sigma \in [0.05, 0.1,\ldots, 1]$ and $\beta \in [0.5, 1,\ldots,10]$, respectively.

In our experiment, we normalize the intensity vector obtained from each local image intensity patch, such that the normalized vector has zero mean and a constant norm, for each label fusion method. Note that the normalization is applied independently at each image location, which may make the resulting voting weights for nearby voxels in an atlas less consistent with each other. To enhance the spatial consistencies of voting weights for nearby voxels, we apply mean filter smoothing with the smoothing window $\mathcal{N}$, the same neighborhood used for local appearance patches, to spatially smooth the voting weights for each atlas.

A. Segmentation of the hippocampus

We use the data in the Alzheimer's Disease Neuroimaging Initiative (ADNI)[3]. ADNI MRI data include 1.5 T structural MRI from all 800 subjects and 3 T structural MRI from 200 subjects. Our study is conducted using only 3 T MRI and only includes data from mild cognitive impairment (MCI) patients and controls. Overall, the data set contains 139 images (57 controls and 82 MCI patients). The images were acquired sagittally, with 1 mm × 1 mm

[3]The ADNI (www.loni.ucla.edu/ADNI) was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year publicprivate partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

in-plane resolution and 1.2 mm slice thickness. To obtain manual segmentation for these data, we first apply a landmark based semi-automatic hippocampal segmentation method [29] to produce the initial segmentation for each image. Each fully-labeled hippocampus was then manually edited using the paintbrush and polygon manual segmentation tools in ITK-SNAP [42] by one of the authors (MA) following a previously validated protocol [15].

Segmentation performance is evaluated using cross-validation. Note that cross-validation is performed twice, once to separate the dataset into an atlas subset and a test subset, and the second time, to search for the optimal value of the label fusion parameters among the atlas subset. For outer cross-validation, we randomly select 20 images to be the atlases and another 20 images for testing. Image-guided registration is performed between all pairs of atlases, and between all atlases and the target image. Global registration was performed using the FSL FLIRT tool [35] with six degrees of freedom and using the default parameters (normalized mutual information similarity metric; search range from −5 to 5 in x, y and z). Deformable registration was performed using the ANTS Symmetric Normalization (SyN) algorithm [4], with the cross-correlation similarity metric (with radius 2) and a Gaussian regularizer with $\sigma = 3$. After registration, reference segmentations from each of the atlases were warped into the target image space.

Fig. 1 illustrates optimal label fusion parameter selection for the three methods in the first cross-validation experiment. The figure plots the number of voxels mislabeled by the automatic segmentation, averaged over 20 inner cross-validation experiments, against the value of each parameter. Note that although the figure plots each parameter separately, the actual search for optimal parameters considers all possible combinations of parameter values. Note that using the appearance window with $r = 1$, all methods performed significantly worse than using larger appearance windows. This indicates that estimation of joint atlas error probabilities in (18) is inaccurate for very small appearance windows. For this cross-validation experiment, the optimal parameters for LWGaussian, LWInverse and LWJoint are ($\sigma = 0.05$, $r=2$, $r_s = 2$), ($\beta = 6$, $r=2$, $r_s = 2$) and ($\beta = 0.5$, $r=2$, $r_s = 3$), respectively.

In all 10 cross-validations, for all three methods the optimal appearance window has radius $r = 2$. Most frequently selected local search windows have radius $r_s = 2$ or $r_s = 3$. For LWGaussian, the most frequently selected weighting model parameters (σ) are 0.05 or 0.1. For LWInverse, the most frequently selected model parameters (β) are located in the range [4,6], while for LWJoint, the most frequently selected β are located in the range [0.5,1.5].

Table I shows the segmentation performance of each method in terms of Dice similarity coefficient between MALF results and reference segmentations. The Dice similarity coefficient is the ratio of the volume of overlap between two segmentations and their average volume [10]. Average overlap (and standard deviation) is reported, with averaging over all 10 outer cross-validation experiments and over the 20 test images in each experiment. In each outer cross-validation experiment, optimal parameters are computed for each method using inner cross-validation. In addition to the local weighted voting methods, we also show the performance by the STAPLE algorithm [40] and majority voting. Overall, STAPLE slightly outperformed majority voting. LWGaussian and LWInverse produced similar results, both significantly outperforming majority voting and STAPLE. LWJoint

outperformed LWGaussian and LWInverse over 1% Dice overlap. This improvement is statistically significant, with $p < 0.00001$ on the paired Student's t-test for each cross-validation experiment. Table I also gives the results produced by each local weighted voting method without using local search. Our method produced significant improvement over the competing methods without local search as well. See Fig. 2 for a segmentation example produced by our method and LWGaussian.

Table II presents the average hippocampal volume in control and MCI cohort obtained using different label fusion techniques. The corresponding Cohen's *d* effect size [14] is also shown (computed as the difference of the sample means of the two cohort, divided by the pooled sample standard deviation). To account for differences in head size, the effect size is computed after normalizing the hippocampal volumes by the subject's intracranial volume. Larger values of Cohen's *d* indicate greater effect, i.e., greater ability to tell cohorts apart based on hippocampal volume. Our method produced more accurate volume measurements than LWGaussian and LWInverse, compared to the reference segmentations. Since volume differences produced by different automatic segmentation methods are all proportional to that of manual segmentation, the hippocampus volume measured using different methods show similar separability between the two population groups. LWJoint yields a slightly better effect size than other MALF methods, indicating the volume measurements produced by LWJoint find a slightly more significant difference between the two populations. All MALF algorithms yield greater effect sizes than manual segmentation, likely due to reduced variance in volume estimation.

B. Hippocampal subfield segmentation

To illustrate the performance of LWJoint on a segmentation problem with multiple labels, we apply it to the problem of automatic segmentation of the subfields of the hippocampal formation from oblique coronal T2-weighted MRI. Our experiments use different similarity weighted voting strategies to improve upon the segmentation results presented in our earlier work [43]. This earlier work also used MALF with spatially varying similarity-weighted label fusion, but the strategy employed there did not consider optimality, and as we show below, performed worse than any of the local weighted voting methods considered here.

1) Imaging Data, Manual Segmentation and Experimental Setup—The experiments use in vivo MRI from 32 subjects from an aging and dementia study [26]. The data were acquired on a Bruker Med-Spec 4T system controlled by a Siemens Trio™ console and equipped with a USA instruments eight channel array coil that consisted of a separate transmit coil enclosing the eight receiver coils. The following sequences, which were part of a larger research imaging and spectroscopy protocol, were acquired: 1. 3D T1-weighted gradient echo MRI (MPRAGE) TR/TE/TI = 2300/3/950 ms, 7° flip angle, $1.0\times1.0\times1.0$ mm^3 resolution, FOV 256×256×176, acquisition time 5.17 min; 2. high resolution T2 weighted fast spin echo sequence (TR/TE: 3990/21 ms, echo train length 15, 18.6 ms echo spacing, 149° flip angle, 100% oversampling in ky direction, 0.4×0.5 mm^2 in plane resolution, 2 mm slice thickness, 24 interleaved slices without gap, acquisition time 3:23 min, oblique coronal slice orientation, angulated perpendicular to the long axis of the hippocampal formation.

The T1-weighted and T2-weighted MR images have complimentary characteristics. The T2-weighted MRIs depict details of the internal structure of the hippocampus with high in-slice resolution and good contrast between subfields, but these images also have a limited field of view and large slice spacing. On the other hand, the T1-weighted images have nearly isotropic voxels and cover the entire brain, but lack contrast between subfields. Manual segmentation protocols that can reliably subdivide the hippocampus into subregions corresponding to its anatomical subfields have been developed for focal T2-weighted data [44], [26]. Our manual segmentation protocol is derived from [26]; it has been expanded to include more slices and additional subfields. Each hippocampus formation is partitioned into anterior (head), posterior (tail), mid-region (body), subiculum (SUB), entorhinal cortex (ERC) and parahippocampal gyrus (PHG). The hippocampal body is further divided into cornu Ammonis fields 1–3 (CA1-3), dentate gyrus (DG), and a miscellaneous label, which contains cysts, arteries, etc. Manual segmentation of hippocampal subfields is unreliable in the head and tail. The boundaries between head, tail and the body regions are defined by a pair of slices in the MRI image. Overall, there are nine subfields defined. See [43] for more detail about the manual subfield segmentation.

As in whole-hippocampus segmentation, we perform a series of 10 cross-validation experiments. In each, 22 subjects are randomly selected as atlases, and the remaining 10 are selected for testing. Registration between all pairs of subjects is performed using SyN using a multi-modality similarity term that assigns equal weight to the T1-weighted image similarity and T2-weighted image similarity (see [43] for details). However, as in [43], similarity-weighted label fusion is applied only using T2-weighted image intensities. The motivation for this is that the T1-weighted MR images mainly serve to align the hippocampal region, while the T2-weighted MR image provide details used for subfield alignment.

Since the in-plane resolution of the T2-weighted images is much higher than slice thickness, instead of the cubic neighborhood definition used in the whole hippocampus segmentation experiment, we use an anisotropic neighborhood definition $2r \times 2r \times 3$ for the local appearance window and $2r_s \times 2r_s \times 3$ for the local searching window in this experiment, with r and r_s specifying the in-plane neighborhood radius. For instance, $r = 3$ and $r_s = 3$ both define a neighborhood of size $7 \times 7 \times 3$.

2) Results—Fig. 3 illustrates some of the parameter selection experiments for the three methods using the atlases in the first outer cross-validation experiment. For this cross-validation experiment, the optimal parameters for LWGaussian, LWInverse and LWJoint are ($\sigma = 0.1$, $r=3$, $r_s = 3$), ($\beta = 5$, $r=3$, $r_s = 3$) and ($\beta = 2$, $r=3$, $r_s = 3$), respectively.

In all 10 cross-validations, all three label fusion methods select the same optimal appearance window and optimal local searching window, which are $r = 3$ and $r_s = 3$, respectively. For LWGaussian, the most frequently selected weighting model parameters (σ) are 0.1 or 0.15. For LWInverse, the most frequently selected model parameters (β) are located in the range [4,6], while for LWJoint, the most frequently selected β are in the range [1.5,2.5].

Table III reports the average segmentation accuracy, relative to manual segmentation, obtained by majority voting, STAPLE[4], our earlier work [43], LWInverse, LWGaussian, and LWJoint. For this application, STAPLE performed comparably to majority voting. Again, LWGaussian and LWInverse produced similar performance, which is significantly better than majority voting and STAPLE. Our method produced the best average Dice similarity for all subfields. On average, we outperformed similarity-based label fusion methods by ~1% Dice similarity. Improvements made by LWJoint for most subfields are statistically significant. Table III also reports the optimal results produced by each local weighted voting methods without applying local search, our joint method outperformed the competing methods in most subfields as well. Fig. 4 shows some segmentation results produced by different methods.

Note that since the same set of cross-validation experiments (with the same atlas/test partitions in each) were conducted in [43], the segmentation accuracy reported there can be directly compared with our results. In [43], a different local weighting algorithm was implemented. Instead of using Gaussian or inverse distance weighting, weights used in [43] were normalized by the image similarity rank. Unlike our implementation of image similarity based local weighted label fusion, the implementation in [43] did not apply the local searching technique. Furthermore, [43] did not select the optimal parameters for each cross-validation experiment. As a result, the Dice scores produced by LWInverse and LWGaussian similarity-based weighted voting without local search are slightly better than those reported in [43], and the results with local search are on average about 1% higher for all subfields. Our method outperformed [43] by about ~2% on average.

The manual segmentation protocol in [43] always uses slice boundaries to delimit the hippocampal body from the head and tail. This is an artificial boundary, necessitated by the anisotropy of the T2 images, for which there is no real anatomical counterpart. In fact, head, body and tail do not truly constitute different anatomical regions, but rather separate regions of the hippocampus where partitioning into subfields is deemed reliable or unreliable, due to the bending and folding of the hippocampus. Thus, to make the comparisons between automatic and manual segmentation more fair, [43] allows the automatic algorithm to make use of the manual partitioning of slices into head, body, and tail. This is accomplished by a heuristic “fix-up” algorithm. For example, if MALF labels a voxel as HEAD, but the slice is considered a body slice by the manual rater, the MALF result is changed by choosing the body subfield (i.e., CA1-4, DG, or SUB) with the highest label probability. Table IV presents the results of LWGaussian, LWInverse and LWJoint label fusion strategies with/without local search, after applying the fix-up algorithm. Again, LWJoint is the top performer, with accuracy for all subfields except TAIL within 1.5% of the inter-rater precision.

[4]The STAPLE algorithm used in this experiment is implemented by CMTK (available at http://www.nitrc.org/projects/cmtk) from [31], which has a special function for multi-label fusion. In our experiment, we applied the command *imagemath* with the option –*mstaple-disputed* with 20 iterations

V. Conclusions and Discussion

We presented a novel formulation to solve the weighted-voting based label fusion problem. Unlike previous label fusion techniques that independently assign voting weights to each atlas, our method takes the dependencies among the atlases into consideration and attempts to directly reduces the expected label error in the combined solution. Provided estimated pairwise dependencies among the atlases, the voting weights can be efficiently solved in a closed form. In our experiments, we estimated the pairwise dependency terms from local image intensities and compared our method with previous label fusion methods in whole hippocampus segmentation and hippocampus subfield segmentation using MR images. For both problems, our method outperformed competing methods.

a) Comparing to the state of the art in hippocampus segmentation

Since the hippocampus segmentation problem has been widely studied, putting our results in the context helps to reveal the significance of our results. Before making a formal comparison, we note that, as pointed out in [7], direct comparisons of quantitative segmentation results across publications are difficult and not always fair due to the inconsistency in the underlying segmentation protocol, the imaging protocol, and the patient population. However, the comparisons carried out below indicate the highly competitive performance achieved by our label fusion technique.

In the recent hippocampus segmentation literature, some of the best reported accuracy results have been obtained using MALF [7], [8], [20]. All three of these best-performing methods are based on independent label fusion with similarity-based local weighting. Collins et al. [7] and Coupe et al. [8] conduct leave-one-out experiments in a data set containing 80 control subjects, i.e. 79 atlases are used in their experiment. They report average Dice overlaps of 0.887 and 0.884, respectively. In contrast, we report average Dice overlap of 0.900±0.020 for control subjects, more than 1% Dice overlap improvement. For patients with MCI, we report Dice overlap of 0.885±0.028. Leung et al. [20] use a template library of 55 images. However, for each image in the library, both the original and its flipped mirror image are used as atlases. Hence, [20] effectively uses 110 atlases for label fusion.

Leung et al. [20] report results in terms of the Jaccard index ($JI(A,B)=\frac{|A\cap B|}{|A\cup B|}$), reporting an average of 0.80±0.03 for the left side hippocampus in 10 control subjects and 0.81±0.04 in 10 MCI patients. Our results for the left side hippocampus, in terms of JI, are 0.826±0.031 for controls and 0.803±0.041 for MCI patients. Overall, we produced results that compare favorably to the state-of-the-art, using significantly fewer atlases.

b) Computational complexity

Comparing to independently assigning voting weights to each atlas, our method requires an additional step of solving the inverse of the pairwise dependence matrix. Since the number of atlases applied in practice is often small, solving the matrix inverse does not substantially increase the computational burden for label fusion. In fact, the most time consuming step is the local searching algorithm. Without the local searching algorithm, for our hippocampus segmentation experiment on ADNI data, our algorithm segments one hippocampus in a few

seconds on a single core 2G HZ CPU using our current Matlab implementation. Regardless, the vast majority of the computational time is spent performing deformable registration between the atlases and the target image, and the cost of label fusion is negligible in comparison.

c) Relation to the STAPLE algorithm

Comparing to the popular expectation-maximization based STAPLE algorithms [40], [31], there are two key differences in our work. First, like other label fusion methods, the STAPLE algorithms also assume that the segmentation errors from different candidate segmentations are independent. Hence, they can not reduce consistent bias in the candidate segmentations. Second, the classic STAPLE algorithms ignore the appearance information in the target image and the atlas images after the registration. This limitation may affect the reliability of the estimated accuracy in each candidate segmentation.

d) Estimation of the pairwise dependency matrix

The label fusion accuracy of our method depends on the accuracy of the estimated pairwise dependencies between atlases. Hence, one natural way to extend our work is to improve the pairwise dependence estimation. Following the common practice, our current method uses the image intensity to estimate the segmentation label relations. Since local image appearance similarity may be unreliable in predicting registration errors, to further improve the performance, one can incorporate prior knowledge that is empirically learned from the atlases to compliment the similarity-based estimation. For example, to estimate the optimal parameters for label fusion method, we applied a leave-one-out strategy on the set of atlases that segments each atlas using the remaining atlases. These leave-one-out experiments also provide the error redundancy produced by each pair of the remaining atlases in the native space of each segmented atlas. By registering and warping each atlas to a common reference space, one can estimate the empirical average error redundancy between any pair of the atlases. The empirical estimation complements the local appearance based estimation and can be combined with the appearance based estimation for segmenting new images. This is a natural direction for future research.

Acknowledgements

We thank Sussane Mueller and Michael Weiner for providing the images used in our hippocampal subfield segmentation experiments. The data collection was supported by the Grant RO1 AG010897 from National Institutes of Health.

The project described was supported by the awards K25 AG027785 and R01 AG037376 from the National Institute On Aging, and grant 10295 from the Penn-Pfizer Alliance. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute On Aging or the National Institutes of Health.

Data collection and sharing for this project was also funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc, F. Hoffman-La Roche, Schering-Plough, Synarc, Inc., and Wyeth, as well as non-profit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to ADNI are facilitated by the Foundation for the National Institutes of Health (http://www.fnih.org/). The grantee

organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129, K01 AG030514, and the Dana Foundation.

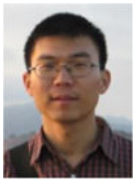
References

1. Aljabar P, Heckemann R, Hammers A, Hajnal J, Rueckert D. Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. NeuroImage. 2009; 46:726–739. [PubMed: 19245840]
2. Allassonniere S, Amit Y, Trouve A. Towards a coherent statistical framework for dense deformable template estimation. Journal of the Royal Statistical Society: Series B. 2007; 69(1):3–29.
3. Artaechevarria X, Munoz-Barrutia A, Ortiz de Solorzano C. Combination strategies in multi-atlas image segmentation: Application to brain MR data. IEEE TMI. 2009; 28(8):1266–1277.
4. Avants B, Epstein C, Grossman M, Gee J. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis. 2008; 12(1):26–41. [PubMed: 17659998]
5. Blezek D, Miller J. Atlas stratification. Medical Image Analysis. 2007; 11(5):443–457. [PubMed: 17765003]
6. Carmichael O, Aizenstein H, Davis S, Becker J, Thompson P, Meltzer C, Liu Y. Atlas-based hippocampus segmentation in Alzheimer's Disease and mild cognitive impairment. NeuroImage. 2005; 27(4):979–990. [PubMed: 15990339]
7. Collins D, Pruessner J. Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion. NeuroImage. 2010; 52(4):1355–1366. [PubMed: 20441794]
8. Coupe P, Manjon JV, Fonov V, Pruessner J, Robles N, Collins DL. Nonlocal patch-based label fusion for hippocampus segmentation. MICCAI. 2010
9. de Leon M, DeSanti S, Zinkowski R, Mehta P, Pratico D, Segal S, Rusinek H, Li J, Tsui W, Louis L, Clark C, Tarshish C, Li Y, Lair L, Javier E, Rich K, Lesbre P, Mosconi L, Reisberg B, Sadowski M, DeBernadis J, Kerkman D, Hampel H, Wahlund L-O, Davies P. Longitudinal CSF and MRI biomarkers improve the diagnosis of mild cognitive impairment. Neurobiol. Aging. 2006; 27(3): 394–401. [PubMed: 16125823]
10. Dice L. Measure of the amount of ecological association between species. Ecology. 1945; 26:297–302.
11. Fischl B, Salat D, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany A, Kennedy A, Klaveness S, Montillo A, Makris N, Rosen B, Dale A. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. Neuron. 2002; 33:341–355. [PubMed: 11832223]
12. Guimond A, Meunier J, Thirion JP. Average brain models: A convergence study. Computer Vision and Image Understanding. 2000; 77(2):192–210.
13. Hansen LK, Salamon P. Neural network ensembles. IEEE Trans. on PAMI. 1990; 12(10):993–1001.
14. Hartung, J.; Knapp, G.; Sinha, BK. Statistical Meta-Analysis with Application. Wiley; 2008.
15. Hasboun D, Chantome M, Zouaoui A, Sahel M, Deladoeuille M, Sourour N, Duymes M, Baulac M, Marsault C, Dormont D. MR determination of hippocampal volume: Comparison of three methods. Am J Neuroradiol. 1996; 17:1091–1098. [PubMed: 8791921]
16. Heckemann R, Hajnal J, Aljabar P, Rueckert D, Hammers A. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage. 2006; 33:115–126. [PubMed: 16860573]
17. Isgum I, Staring M, Rutten A, Prokop M, Viergever M, van Ginneken B. Multi-atlas-based segmentation with local decision fusionapplication to cardiac and aortic segmentation in CT scans. IEEE Trans. on MI. 2009; 28(7):1000–1010.
18. Joshi S, Davis B, Jomier M, Gerig G. Unbiased diffeomorphism atlas construction for computational anatomy. NeuroImage. 2004; 23:151–160.

19. Kittler J. Combining classifiers: A theoretical framework. Pattern Analysis and Application. 1998; 1:18–27.
20. Leung K, Barnes J, Ridgway G, Bartlett J, Clarkson M, Macdonald K, Schuff N, Fox N, Ourselin S. Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's Disease. NeuroImage. 2010; 51:1345–1359. [PubMed: 20230901]
21. Liu C, Yuen J, Torralba A. Nonparametric scene parsing: Label transfer via dense scene alignment. CVPR. 2009
22. Lotjonen JM, 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]
23. Malykhin N, Lebel R, Coupland N, Wilman A, Carter R. In vivo quantification of hippocampal subfields using 4.7 T fast spin echo imaging. NeoroImage. 2009; 49(2):1224–1230.
24. Morra J, Tu Z, Apostolova L, Green A, Toga A, Thompson P. Automatic subcortical segmentation using a contextual model. MICCAI. 2008:194–201. [PubMed: 18979748]
25. Mueller S, Stables L, Du A, Schuff N, Truran D, Cashdollar N, Weiner M. Measurements of hippocampal subfields and age related changes with high resolution MRI at 4T. Neurobiology of Aging. 2007; 28(5):719–726. [PubMed: 16713659]
26. Mueller S, Weiner M. Selective effect of age, Apo e4, and Alzheimer's Disease on hippocampal subfields. Hippocampus. 2009; 19(6):558–564. [PubMed: 19405132]
27. Murty, KG. Linear Complementarity, Linear and Nonlinear Programming. Helderman-Verlag; 1988.
28. Penrose L. The elementary statistics of majority voting. Journal of the Royal Statistical Society. 1946; 109(1):53–57.
29. Pluta J, Avants B, Glynn S, Awate S, Gee J, Detre J. Appearance and incomplete label matching for diffeomorphic template based hippocampus segmentation. Hippocampus. 2009; 19:565–571. [PubMed: 19437413]
30. Rohlfing T, Brandt R, Menzel R, Maurer C. 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]
31. Rohlfing T, Russakoff D, Maurer C. Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. IEEE Trans. on Medical Imaging. 2004; 23(8):983–994. [PubMed: 15338732]
32. Sabuncu M, Balci S, Shenton M, Golland P. Image driven population analysis through mixture-modeling. IEEE Trans. on MI. 2009; 28(9):1473–1487.
33. Sabuncu M, Yeo BTT, Van Leemput K, Fischl B, Golland P. A generative model for image segmentation based on label fusion. IEEE TMI. 2010; 29(10):1714–1720.
34. Scahill R, Schott J, Stevens J, Rossor M, Fox N. Mapping the evolution of regional atrophy in Alzheimer's Disease: unbiased analysis of fluidregistered serial MRI. Proc. Natl. Acad. Sci. U. S. A. 2002; 99(7):4703–4707. [PubMed: 11930016]
35. Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, JohansenBerg H, Bannister PR, Luca MD, Drobnjak I, Flitney DE, Niazy RK, Saunders J, Vickers J, Zhang Y, Stefano ND, Brady JM, Matthews PM. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage. 2004; 23(Suppl 1):S208S219.
36. Squire L. Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. Psychological Review. 1992; 99:195–231. [PubMed: 1594723]
37. Toyoda T, Hasegawa O. Random field model for integration of local information and global information. IEEE Trans. on Pattern Analysis and Machine Intelligence. 2008; 30(8):1483–1489.
38. Wang, H.; Suh, J.; Pluta, J.; Altinay, M.; Yushkevich, P. Information Processing in Medical Imaging, volume 6801 of Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2011. Optimal weights for multi-atlas label fusion. In Gbor Szkely and HorstK. Hahn, editors; p. 73-84.
39. Wang H, Suh JW, Das S, Pluta J, Altinay M, Yushkevich P. Regression-based label fusion for multi-atlas segmentation. CVPR. 2011

40. Warfield S, Zou K, Wells W. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE TMI. 2004; 23(7):903–921.
41. Wolz R, Aljabar P, Hajnal JV, Hammers A, Rueckert D. Leap: Learning embeddings for atlas propagation. NeuroImage. 2010; 49(2):1316–1325. [PubMed: 19815080]
42. Yushkevich P, Piven J, Hazlett H, Smith R, Ho S, Gee J, Gerig G. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. NeuroImage. 2006; 31(3):1116–1128. [PubMed: 16545965]
43. Yushkevich P, Wang H, Pluta J, Das S, Craige C, Avants B, Weiner M, Mueller S. Nearly automatic segmentation of hippocampal subfields in in vivo focal T2-weighted MRI. NeuroImage. 2010; 53(4):1208–1224. [PubMed: 20600984]
44. Zeineh M, Engel S, Thompson P, Bookheimer S. Dynamics of the hippocampus during encoding and retrieval of face-name pairs. Science. 2003; 299(5606):577–580. [PubMed: 12543980]

Biographies



Hongzhi Wang received his Bachelor, Master degrees in computer science from University of Science and Technology Beijing in 2000 and 2003, respectively. In 2008, he received his PhD degree in computer science from Stevens Institute of Technology. His PhD research focused on perceptual organization and its application in shape-based object recognition. Since then he has been a postdoctoral researcher at Penn Image Computing and Science Lab, University of Pennsylvania, working on applications in medical image analysis. His research interests include developing techniques that allow effective information fusion for accurate quantitative measurements from medical images. Dr. Wang is a member of the IEEE.

Jung W. Suh received B.S. and M.S. degrees in electronics and computer engineering from Hanyang University, Korea, in 1996, 1998 and a Ph.D. degree in electrical and computer engineering from the Virginia Tech, Blacksburg, in 2007. From 1998 through 2003, he was with the Samsung Electronics Company, Korea, where he was involved in the development of MPEG-4 and Digital Mobile Broadcasting (DMB) systems. He is currently a senior medical imaging scientist at HeartFlow, Inc. His research interests are in the fields of biomedical image processing, computer vision and pattern recognition, image enhancement and compression.

Sandhitsu Das received his B.Tech. M.Tech. degrees in Electrical Engineering from the Indian Institute of Technology, Kanpur in 1997 and 1999, respectively. He obtained his Ph.D. in Bioengineering, studying computational models of human visual system, from the School of Engineering and Applied Sciences at the University of Pennsylvania in 2006. Since then, he has been at the Penn Image Computing and Science Laboratory at the Department of Radiology, where he currently holds the position of Senior Research Investigator. His research interests include developing novel techniques for analyzing structural and functional brain imaging data, and their applications to clinical and basic neuroscience. His current work includes development of imaging-based biomarkers for various brain disorders like epilepsy, Alzheimer's disease, and other forms of dementia, as well as measurements for studying mechanisms of human episodic memory using functional MRI.

John B. Pluta was born in Chester, Pennsylvania, in 1982. He attended Drexel University in Philadelphia, Pennsylvania, from 2000–2005 and graduated with a B.S. in psychology. He is currently a graduate student at the University of Pennsylvania, studying biostatistics.

He began his career in 2004, as a research assistant at Moss Rehabilitation Institute. After graduating, he took a position as a Research Specialist at the University of Pennsylvania. His main areas of interest are hippocampal anatomy, segmentation methods, and image registration.

Caryne Craige was born in Philadelphia, PA on June 20, 1984. Ms. Craige achieved a Bachelor of Science degree in biology in 2006 from the University of Richmond in Richmond, VA, U.S.A. She worked as a Research Technician at the Children's Hospital of Philadelphia from March 2007-June 2009, working within the Stress Neurobiology group, in the laboratory of Dr. Sheryl Beck. With this position, Ms. Craige studied dorsal raphe and hippocampal neurocircuitry in the context of stress disorders, with the use of transgenic and behavioral animal models. In January of 2008, Ms. Craige began work as a Research Assistant in the laboratory of Dr. Paul Yushkevich at the University of Pennsylvania. Her

work in this laboratory is ongoing and her main contributions entail hippocampal subregion manual segmentation of T2- and T4-weighted magnetic resonance images using ITK-SNAP software. Ms. Craige is currently affiliated with Temple University School of Medicine in Philadelphia, PA, U.S.A., where she is a third-year graduate student in the Pharmacology Ph.D. program. Ms. Craige is a member of the Society for Neuroscience.

Paul A. Yushkevich (M06) became a Member (M) of IEEE in 2006. He was born in Moscow, Russia. He holds the degrees of B.S. in Computer Science and Mathematics (1996) from the University of North Carolina at Charlotte; M.S. in Computer Science (2000) and Ph.D. in Computer Science (2003) from the University of North Carolina at Chapel Hill.

He is an Assistant Professor of Radiology at the University of Pennsylvania, Philadelphia, PA, USA since 2009. Previously, he was a Research Assistant Professor (2006–2009) and a Postdoctoral Fellow (2003–2008) at the University of Pennsylvania. He has authored 30 journal publications and 34 peer-reviewed conference papers, spanning many areas of the biomedical image analysis field, including geometrical object representation, shape analysis, image segmentation, and image registration. His applied research primarily focuses on the problem of developing imaging biomarkers in aging and dementia, but also includes other neuroimaging applications and cardiac imaging analysis. He maintains and develops ITK-SNAP software for interactive image segmentation (itksnap.org), which is among the most widely used open-source image analysis tools. Prof. Yushkevich is a Member of the IEEE BISP.

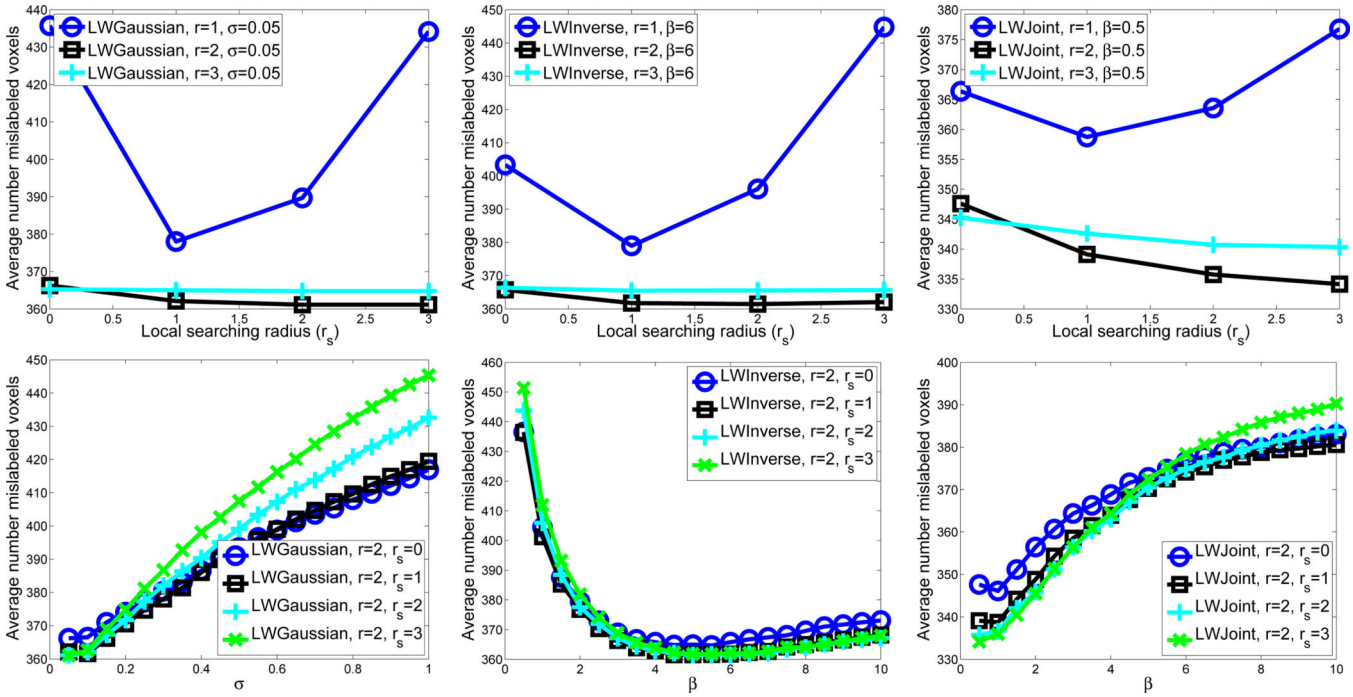


Fig. 1.

Optimal label fusion parameter selection for LWGaussian (left), LWInverse (middle) and LWJoint (right) using leave-one-out cross-validation. The upper figures plot the average number of mislabeled voxels against the local searching radius r_s and the appearance window radius r. The weighting function parameters σ, β are held fixed in these figures at its optimal value for the three methods, respectively. The lower figures plot the average number of mislabeled voxels against the local searching radius r_s and the weighting function parameter, σ and β, respectively. The appearance window radius r is held fixed in this figure at its optimal value.

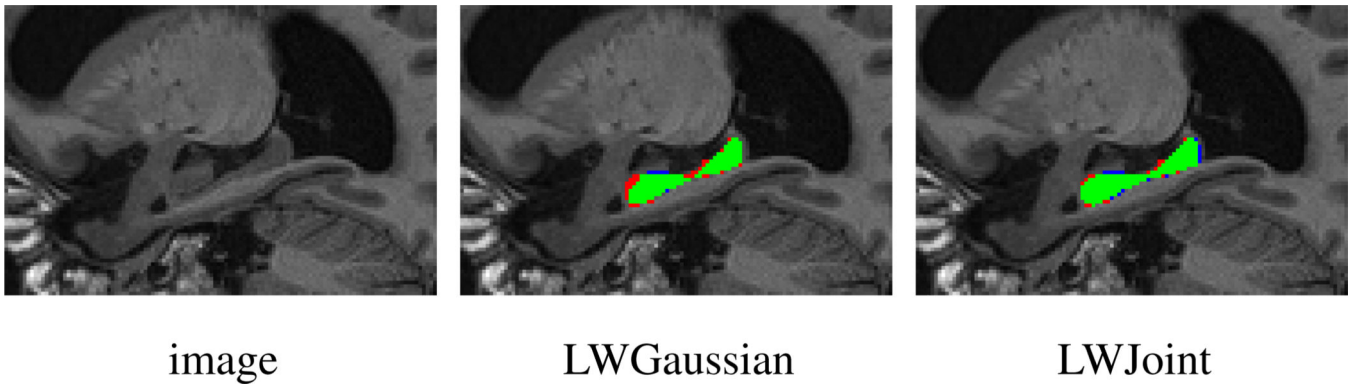


Fig. 2.

Sagittal views of a segmentation produced by LWGaussian and our method. Red: reference segmentation; Blue: automatic segmentation; Green: overlap between manual and automatic segmentation.

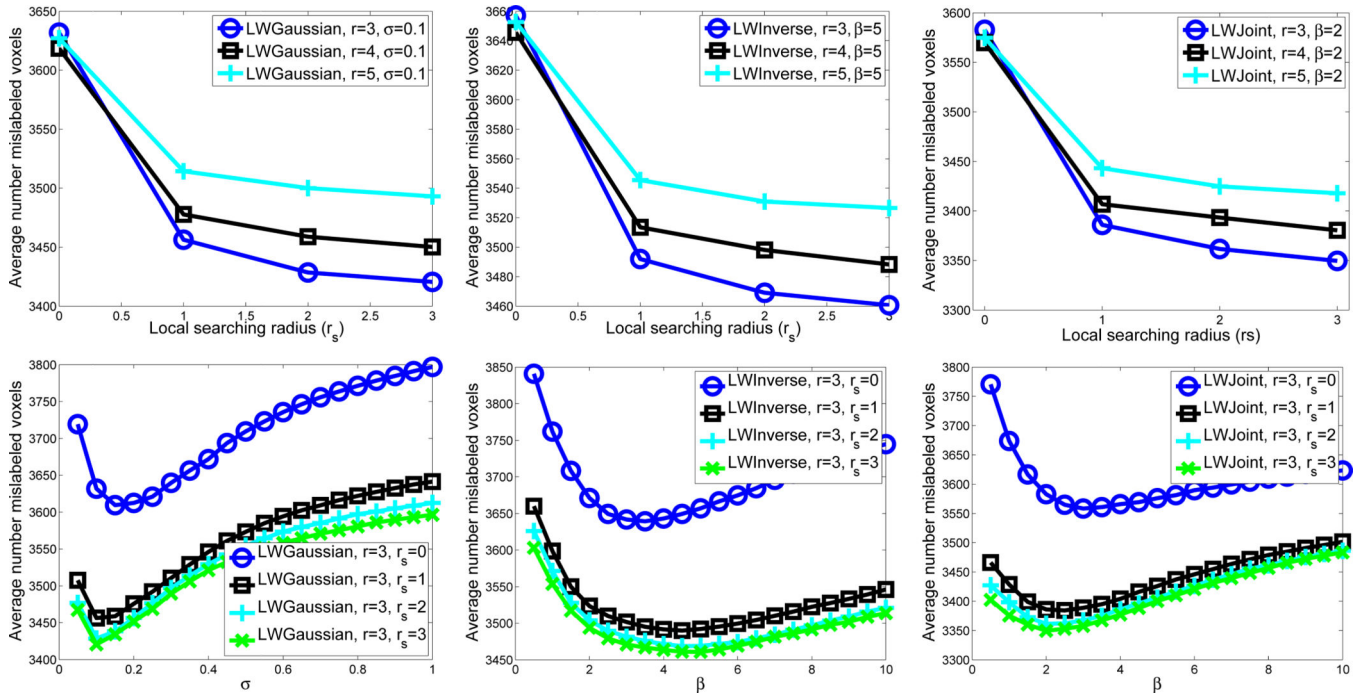


Fig. 3.

Optimal label fusion parameter selection for LWGaussian (left), LWInverse (middle) and LWJoint (right) using leave-one-out cross-validation. The upper figures plot the average number of mislabeled voxels against the local searching radius r_s and the appearance window radius r. The weighting function parameters σ, β are held fixed in these figures at its optimal value for the three methods, respectively. The lower figures plot the average number of mislabeled voxels against the local searching radius r_s and the weighting function parameter, σ and β, respectively. The appearance window radius r is held fixed in this figure at its optimal value.

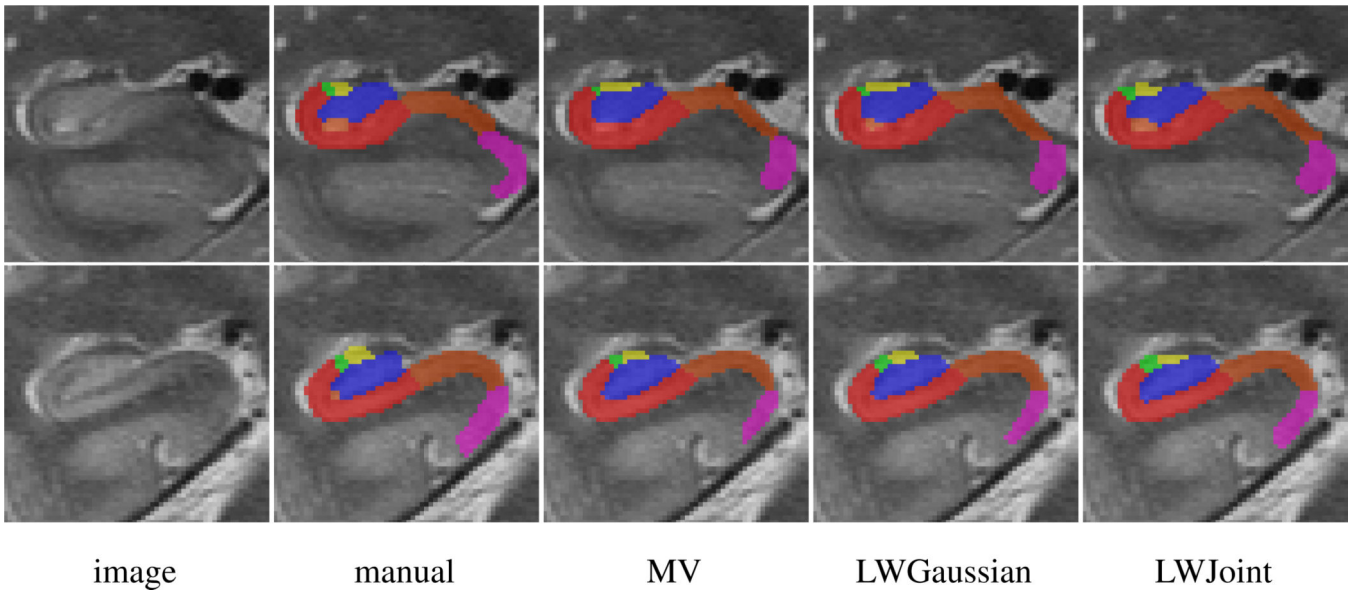


Fig. 4.

Coronal views of some subfield segmentation results produced by MV, LWGaussian and our method. All results are produced with local searching using the optimal parameter for each method. Label description: red - CA1; green - CA2; yellow - CA3; blue - DG; light brown - miscellaneous label; brown - SUB; cyan - ERC; pink - PHG.

TABLE I

whole hippocampus segmentation performance for each label fusion method, in terms of Dice similarity between MALF results and reference segmentations. Average dice similarity (± standard deviation) are computed across 10 outer cross-validation experiments, each having 20 test images. The results produced by each local weighted voting methods without applying local search are shown in parenthesis. Greatest similarity is obtained using the proposed LWJoint method.

Label Fusion Strategy	Dice Similarity (Left Hippocampus)	Dice Similarity (Right Hippocampus)
Majority Voting	0.836 ± 0.084	0.829 ± 0.069
STAPLE [40]	0.846 ± 0.086	0.841 ± 0.086
LWGaussian	(0.885 ± 0.025) 0.886 ± 0.027	(0.873 ± 0.030) 0.875 ± 0.030
LWInverse	(0.884 ± 0.026) 0.885 ± 0.027	(0.872 ± 0.030) 0.873 ± 0.030
LWJoint	(0.893 ± 0.025) **0.897** ± 0.024	(0.884 ± 0.027) **0.888** ± 0.026

TABLE II

Hippocampal volume (mm^3) (left/right) measured by different label fusion methods for control and MCI cohorts. The results are averaged over 10 cross-validation experiments, which together include test images from 94 control subjects and 106 MCI subjects. The last column shows the corresponding Cohen's *d* effect size, whose magnitude indicates the difference between the two populations. The Hippocampus volume is normalized by intracranial volume for computing the Cohen's effect size.

	Left Hippocampus			Right Hippocampus		
Label Fusion Method	**Volume (CTL)**	**Volume (MCI)**	**Cohen's *d***	**Volume (CTL)**	**Volume (MCI)**	**Cohen's *d***
LWGaussian	2026 ± 277	1642 ± 334	1.7260	1947 ± 311	1553 ± 346	1.5576
LWInverse	2014 ± 274	1635 ± 326	1.7266	1930 ± 309	1544 ± 338	1.5504
LWJoint	2156 ± 285	1755 ± 353	**1.7468**	2083 ± 322	1668 ± 373	**1.5700**
Reference Seg.	2258 ± 325	1841 ± 368	1.5747	2201 ± 378	1785 ± 408	1.3643

IEEE Trans Pattern Anal Mach Intell. Author manuscript; available in PMC 2013 December 22.

TABLE III

Average performance of different label fusion strategies in hippocampal subfield segmentation experiments. Average Dice similarity (± standard deviation) are shown between the MALF result and corresponding manual segmentation for each subfield, also averaging over left and right hemispheres and over 10 cross-validation experiments. the results produced by each local weighted voting methods without applying local search are shown in parenthesis. Asterisks in the last column indicate results where the improvement of LWJoint over LWGaussian and LWInverse was statistically significant, as per paired student's *t*-test.

Subfield	Maj. Voting	STAPLE [31]	[43] (Table 6)	LWInverse	LWGaussian	LWJoint
CA1	0.731±0.070	0.719±0.102	0.770±0.065	(0.766±0.064) 0.782±0.061	(0.769±0.063) 0.786±0.060	(0.773±0.059) **0.789**±0.057**
CA2	0.322±0.180	0.357±0.191	0.422±0.175	(0.427±0.173) 0.427±0.177	(0.431±0.174) 0.431±0.178	(0.442±0.170) **0.455**±0.176***
CA3	0.497±0.142	0.523±0.141	0.532±0.137	(0.537±0.124) 0.559±0.124	(0.541±0.125) 0.560±0.125	(0.549±0.121) **0.572**±0.123***
DG	0.741±0.090	0.738±0.100	0.773±0.067	(0.768±0.070) 0.785±0.068	(0.767±0.070) 0.786±0.069	(0.767±0.070) **0.789**±0.068*
head	0.864±0.028	0.860±0.033	0.874±0.025	(0.874±0.023) 0.882±0.024	(0.875±0.023) 0.883±0.024	(0.876±0.022) **0.885**±0.023***
tail	0.739±0.123	0.720±0.144	0.744±0.119	(0.743±0.124) 0.752±0.128	(0.746±0.123) 0.755±0.129	(0.748±0.121) **0.759**±0.125**
SUB	0.706±0.062	0.705±0.083	0.727±0.061	(0.726±0.067) 0.733±0.064	(0.729±0.066) 0.736±0.062	(0.734±0.061) **0.745**±0.060***
ERC	0.606±0.138	0.603±0.131	0.627±0.123	(0.630±0.124) 0.647±0.128	(0.632±0.125) 0.648±0.129	(0.634±0.120) **0.652**±0.126*
PHG	0.604±0.076	0.626±0.080	0.625±0.076	(0.629±0.075) 0.632±0.078	(0.629±0.076) 0.632±0.078	(0.630±0.074) **0.640**±0.076***

* $p < 0.05$

** $p < 0.01$

*** $p < 0.0001$

IEEE Trans Pattern Anal Mach Intell. Author manuscript; available in PMC 2013 December 22.

TABLE IV

Average performance of different label fusion strategies in hippocampal subfield segmentation experiments, when using manual head/body/tail slice partitioning (see text). average dice similarity (± standard deviation) are shown between the MALF result and corresponding manual segmentation for each subfield, also averaging over left and right hemispheres and over 10 cross-validation experiments. The results produced by each local weighted voting methods without applying local search are shown in parenthesis. Asterisks in the LWJoint column indicate results where the improvement of LWJoint over LWGaussian and LWInverse was statistically significant, as per paired student's *t*-test. the last column gives average dice overlap between manual segmentations produced by two trained human raters.

Subfield	Maj. Voting	[43] (Table 1)	LWInverse	LWGaussian	LWJoint	Inter-Rater
CA1	0.804±0.059	0.851±0.040	(0.853±0.040) 0.869±0.038	(0.855±0.038) 0.870±0.037	(0.858±0.037) **0.874**±0.035***	0.883±0.032
CA2	0.357±0.194	0.470±0.179	(0.474±0.183) 0.486±0.180	(0.477±0.182) 0.489±0.180	(0.487±0.177) **0.510**±0.176***	0.522±0.160
CA3	0.530±0.145	0.583±0.133	(0.591±0.124) 0.619±0.123	(0.595±0.125) 0.620±0.123	(0.605±0.121) **0.634**±0.123***	0.668±0.087
DG	0.813±0.087	0.859±0.045	(0.855±0.047) 0.870±0.046	(0.857±0.045) 0.871±0.045	(0.858±0.045) **0.875**±0.044***	0.885±0.034
head	0.878±0.021	0.893±0.018	(0.892±0.016) 0.900±0.017	(0.893±0.015) 0.900±0.016	(0.894±0.016) **0.903**±0.016***	0.900±0.016
tail	0.793±0.112	**0.828**±0.105	(0.804±0.105) 0.812±0.108	(0.806±0.104) 0.814±0.109	(0.809±0.103) 0.819±0.105***	0.901±0.059
SUB	0.715±0.062	0.742±0.063	(0.741±0.067) 0.747±0.065	(0.743±0.066) 0.749±0.064	(0.748±0.062) **0.758**±0.062***	0.768±0.079
ERC	0.606±0.138	0.738±0.093	(0.745±0.096) 0.759±0.095	(0.745±0.096) 0.759±0.095	(0.750±0.094) **0.768**±0.090***	0.786±0.123
PHG	0.627±0.074	0.658±0.073	(0.662±0.071) 0.666±0.075	(0.662±0.072) 0.666±0.075	(0.664±0.071) **0.675**±0.072***	0.706±0.106

* $p < 0.05$

** $p < 0.01$

*** $p < 0.0001$