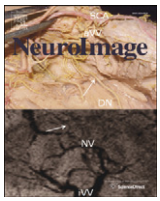
NeuroImage 56 (2011) 126–139

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg



Optimal weights for local multi-atlas fusion using supervised learning and dynamic information (SuperDyn): Validation on hippocampus segmentation

Ali R. Khan [a], Nicolas Cherbuin [b], Wei Wen [c], Kaarin J. Anstey [b], Perminder Sachdev [c], Mirza Faisal Beg [a,*]

[a] School of Engineering Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada
[b] Centre for Mental Health Research, Australian National University, Canberra, ACT 0200, Australia
[c] School of Psychiatry, University of New South Wales, Sydney, NSW 2052, Australia

ARTICLE INFO

Article history:
Received 13 November 2010
Revised 26 January 2011
Accepted 28 January 2011
Available online 4 February 2011

Keywords:
Segmentation
MRI
Hippocampus
Weighted voting
Classifier combination
Segmentation validation

ABSTRACT

We developed a novel method for spatially-local selection of atlas-weights in multi-atlas segmentation that combines supervised learning on a training set and dynamic information in the form of local registration accuracy estimates (SuperDyn). Supervised learning was applied using a jackknife learning approach and the methods were evaluated using leave-N-out cross-validation. We applied our segmentation method to hippocampal segmentation in 1.5T and 3T MRI from two datasets: 69 healthy middle-aged subjects (aged 44–49) and 37 healthy and cognitively-impaired elderly subjects (aged 72–84). Mean Dice overlap scores (left hippocampus, right hippocampus) of (83.3, 83.2) and (85.1, 85.3) from the respective datasets were found to be significantly higher than those obtained via equally-weighted fusion, STAPLE, and dynamic fusion. In addition to global surface distance and volume metrics, we also investigated accuracy at a spatially-local scale using a surface-based segmentation performance assessment method (SurfSPA), which generates cohort-specific maps of segmentation accuracy quantified by inward or outward displacement relative to the manual segmentations. These measurements indicated greater agreement with manual segmentation and lower variability for the proposed segmentation method, as compared to equally-weighted fusion.

Introduction

Accurate and robust segmentation of subcortical and neocortical structures in T1-weighted magnetic resonance images (MRI) has many useful and important applications in morphometric studies of disease progression and aging. One such application, shape analysis, allows for quantification of a brain structure's shape features, capturing the subtle and local structural changes that can accompany disease or abnormal development. Many shape analysis methods rely on segmentation to enable comparison of shape cross-sectionally over a population, or longitudinally over time. Segmentation methods that can consistently and precisely delineate the structural boundaries can potentially generate more meaningful results and better detect the minute changes associated with disease progression. Shape analysis of the hippocampus, a medio-temporal lobe structure known for its involvement in long-term memory, has been applied in numerous studies of schizophrenia (Styner et al., 2004; Wang et al., 2008), Alzheimer's disease (Wang et al., 2003; Thompson et al., 2004; Wang et al., 2006; Apostolova et al., 2006; Tepest et al., 2008; Wang et al., 2009), and temporal lobe epilepsy (Hammers et al., 2007). Thus, improved accuracy and reliability of hippocampal segmentation is a relevant issue for future morphometric studies, if methods are to be eventually used for clinical purposes.

Manual tracing is widely considered to be closest to gold standard in hippocampal segmentation, but possesses several drawbacks when used in large studies, including the significant time-cost of trained raters and issues of intra- and inter-rater reliability. Therefore, automated segmentation methods are becoming more desirable as the size and availability of large MRI databases increases. Atlas-based label propagation is a simple method for automated segmentation that involves performing non-rigid registration between a labeled atlas image and a target, propagating the labels from the atlas to the target to generate a labeling for the target (Bajcsy et al., 1983). The accuracy of the resulting segmentation thus depends on the ability of the registration to find accurate and meaningful correspondence between the images, which is inherently related to the anatomical similarity of the two images. This therefore brings up the question of atlas selection, since using an atlas that is anatomically similar to a given subject would result in better performance for that subject, but would possibly sacrifice performance on additional different subjects. A better choice then becomes the atlas that can best segment the majority of images, but because of the inherent anatomical variability present in most datasets, it would likely under-perform on the outlying images.

Multi-atlas segmentation attempts to resolve this bias by using a number of different subjects as atlas images, performing multiple registrations from all the atlases to the target and fusing the results to

* Corresponding author.
E-mail addresses: akhanf@sfu.ca (A.R. Khan), mfbeg@ensc.sfu.ca (M.F. Beg).

1053-8119/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2011.01.078

generate the target segmentation (Rohlfing et al., 2004; Heckemann et al., 2006). Label fusion is commonly achieved using a majority vote rule. Using multiple atlases in this way produces better results because the anatomical variability is represented more accurately than in a single atlas, and errors due to mis-labeling or mis-registration are averaged out when the individual propagated labels are fused together. Template selection in this case still remains an issue, however now the problem becomes choosing the multiple atlases, rather than picking a single atlas subject. Furthermore, in both cases, atlas selection for one region of the image may not be ideal for another region of the image, thus, incorporating spatial locality in the atlas selection is also preferred.

Several recent methods have approached the problem of atlas selection in multi-atlas segmentation (Isgum et al., 2009; Artaechevarria et al., 2009; van Rikxoort et al., 2010; Sabuncu et al., 2010; Aljabar et al., 2009; Sdika, 2010; Leung et al., 2010). Borrowing terminology from the analogous field of classifier combination, these methods can be partitioned into those that determine the selection weights *a priori* (static selection), and those that determine appropriate selection weights for each new image (dynamic selection).

The segmentation accuracy map weights of (Sdika, 2010) are an example of static selection, since these represent the local accuracy of each atlas as inferred from segmentation of a labeled training set with each atlas. Selection methods using supervised learning are advantageous because they provide a set of weights that perform optimally well under the given conditions. However, these methods may suffer from over-fitting to the training set, and thus may perform poorly if the target images deviate considerably from the training set. The usual requirement of a labeled training set is also a constraint that may be difficult to meet considering the potentially high cost of manual segmentation. Recent work in classifier combination (Ruta and Gabrys, 2005) has also shown that team performance should be used instead of individual classifier performance when selecting classifiers, thus methods that set atlas weights concurrently instead of one atlas at a time (Sdika, 2010) should be desirable.

One example of dynamic selection is the use of atlas-target registration accuracy estimators to weight the influence of a given atlas (Isgum et al., 2009; Artaechevarria et al., 2009; van Rikxoort et al., 2010; Wu et al., 2007), based on the presumption that poor registration will imply a poor segmentation, and thus the corresponding atlas should be weighted less. Similarly, methods which employ demographics or image similarity metrics, such as mutual information, to select atlases (Aljabar et al., 2009) are also examples of dynamic selection, which presume that choosing atlases that are similar to the target will result in more accurate segmentations. The use of dynamic information is advantageous because it has the potential for applicability over a wider range of images, since the inherent properties of each target image and its relationship to the atlases are considered. However, the metrics used to assess the dynamic information, such as registration accuracy or image similarity, are mainly heuristic in nature, and so the actual relationship to the atlas weights is generally not known.

We propose a method that employs a combination of both supervised learning (static selection) and dynamic selection to determine the spatially-local optimal weights for a given set of atlases (SuperDyn). Local estimates of registration accuracy are used as predictors, combined jointly with the propagated labels, in a multiple linear regression to find the weight coefficients that leads to the best-fit to the training labels. Thus, applied to a target subject, the local estimates of registration accuracy (dynamic selection) and the weight coefficients from supervised learning (static selection) form an effective atlas weight to be used in weighted segmentation fusion.

In addition to the segmentation methodology itself, validation is an issue that deserves much attention. Segmentation methods are commonly validated using global segmentation similarity metrics, such as volumetric overlap, surface distance, and volume differences or correlations. These describe the overall similarity of one segmentation to another but they do not describe the performance of the segmentation locally and do not convey any information about local variability of the segmentations. We describe a method for local surface-based segmentation performance assessment (SurfSPA) that describes the local inward or outward displacements of an automated segmentation, relative to the manual segmentation, and allows for cohort-based statistics of these measures, including the mean or variance of these surface–surface displacements.

In this paper we will first introduce the methodology of weighted segmentation fusion, using supervised, dynamic, and supervised + dynamic weights, and describe how a regularized least-squares solution analogous to multiple linear regression was found. A jackknife-based learning method will be described to learn weights given a single set of labeled images. An overview of the existing label propagation method will then be given, and we will describe how registration accuracy estimates are used as dynamic information. The segmentation performance metrics, both global and local will be described, and we will show results of hippocampus segmentation experiments using cross-validation on a two datasets of 69 and 37 subjects, comparing our SuperDyn method to supervised-only, dynamic-only, equal weighted fusion, and STAPLE.

Methods

In this section we will describe the general idea of optimally-weighted segmentation fusion, using regularized supervised learning and dynamic information to determine the weights, and a jackknife learning algorithm to learn optimal weights for a single group of labeled images. We will then describe how the proposed methods were applied to hippocampus segmentation using FS + LDDMM (Freesurfer + large deformation diffeomorphic metric mapping) multi-atlas label propagation, using a local estimate of registration accuracy for dynamic selection.

First, we introduce the notation used in describing the optimally-weighted segmentation fusion. Anatomical images are represented as functions, $I:\Omega\rightarrow\mathbb{R}$, where Ω is the three-dimensional image domain, and the corresponding image intensities at each voxel are represented by real numbers. In multi-atlas label propagation, we denote the target images as I_i and the M atlas images as $I_j, j=1\ldots M$. Segmentations of given structures, $S:\Omega\rightarrow[-1,+1]\in\mathbb{R}$ are represented as continuous images with a maximum value of $+1$ for the foreground, and a minimum value of -1 for the background, by applying a normalization function of $q=\frac{2p-P}{P}$, where the input segmentation intensity is p and has a minimum and maximum of 0 and P respectively. Specifically, the manual segmentation of the i-th image is denoted as S_i^{man}, and the automated segmentation of this image obtained using the j-th atlas is $S_{i,j}^{auto}$.

Locally-weighted segmentation fusion

Let us consider the case of weighted fusion for a target image, I_i at a single voxel location, $x\in\Omega$. Given a set of M labeled atlases, I_j and S_j^{man}, $j=1\cdots M$, we can perform atlas-based segmentation to obtain M segmentations, $S_{i,j}^{auto}$. If we assume M corresponding weights are also known, $\beta_j(x)$, $j=1\ldots M$, which are larger in magnitude when the segmentation is expected to be closer to the true value, then the weighted fusion $S^{weighted}$ at each voxel can be written as

$$S_i^{weighted}(x)=\sum_{j=1}^{M}\beta_j(x)S_{i,j}^{auto}(x).$$

When β_j is zero, its corresponding segmentation, $S_{i,j}^{auto}$ is effectively unused in the fused segmentation, since it would contribute exactly zero to the fused segmentation. And conversely, if $|\beta_j|$ is greater than

zero, the corresponding segmentation would effectively contribute an amount proportional to the size of β_j. Note that majority voting with discrete labels is equivalent to equal-weighted fusion, S_i^{Eql} with $\beta_j(x) = frac\frac{1}{M}, \forall j$, followed by thresholding of the segmentation at 0 (since segmentations are represented in $[-1,+1] \in \mathbb{R}$). Our choice of a linear model for segmentation fusion should be sufficient for complex structures like the hippocampus since we incorporate spatially-local and dynamic weighting, which allows for different treatment of adjacent regions.

Supervised weight selection

Now, consider the case where ground truth, or manual segmentations, are given for a set of N training images, D_{train}^N and we would like to determine the optimal atlas weights for the M atlases. Note that when dealing with multiple training subjects, we have to ensure they are spatially normalized to account for gross anatomical variability, so that voxel-wise comparisons can be made. We use a local affine transformation to perform this normalization, which is described in more detail in Spatial normalization section.

Using the weighted fusion model as before, at each voxel, x, we can write a set of N equations, $e_i(x), i = 1 \ldots N$, one for each training sample, which we can solve for $M+1$ unknowns, $\beta_j(x), j = 0 \ldots M$ (including a bias parameter $\beta_0(x)$):

$$e_i(x) : S_i^{man}(x) = \beta_0(x) + \sum_{j=1}^{M} \beta_j(x) S_{i,j}^{auto}(x); i = 1 \ldots N.$$

If we have fewer atlas subjects than training subjects, then we have an overdetermined linear system ($M<N$) that can be solved via a least-squares approach. The design matrix at each voxel, denoted by Φ_x, is an $N \times (M+1)$ matrix with each $(i,j)^{th}$ entry equal to $S_{i,j}^{auto}(x), i = 1 \ldots N$, $j = 0 \ldots M$, and with $S_{i,j=0}^{auto} = 1$ for the bias term. Let $\mathbf{s}_x^{man}$ represent the column vector of training manual segmentations at voxel x, and let β_x represent the column vector of weights at voxel x. The linear system can be written as $\Phi_x \beta_x = \mathbf{s}_x^{man}$, then the linear least-squares solution is found by minimizing the residual $\| \Phi_x \beta_x - \mathbf{s}_x^{man} \|^2$. Finally, the least-squares estimate of the weights can then be written as:

$$\beta_x = \Phi_x^{\dagger} \mathbf{s}_x^{man},$$

where $\Phi_x^{\dagger} = (\Phi_x^T \Phi_x)^{-1} \Phi_x^T$ is simply the pseudo-inverse of Φ_x. Note that this approach is analogous to multiple linear regression. Once the atlas weights have been found, an additional image, $I_\star$ can then be segmented using the weighted fusion of its atlas segmentations at each location, x,

$$S_\star^{Super}(x) = \sum_{j=0}^{M} \beta_j(x) S_{\star,j}^{auto}(x); S_{\star,0}^{auto}(x) = 1.$$

Here, S^{Super} is used to denote the fused segmentation using weights obtained from supervised learning on the training set.

Regularization

This linear least squares solution in practice can become problematic, since training data is generally scarce because of the time-cost of manually-traced segmentations. Thus, if the number of parameters (effectively the number of atlases used, M) is relatively high compared to the number of training subjects, over-fitting can occur. Regularization can be used to address this problem, which we implement in two forms: Tikhonov regularization and a spatial regularization scheme we refer to as spatial-neighborhood sampling.

In Tikhonov regularization, a regularization term is included in the least-squares minimization in order to give preference to solutions with small norms, leading to a minimization of $\| \Phi_x \beta_x - \mathbf{s}_x^{man} \|^2 + \lambda \| \beta_x \|^2$, where λ is a free parameter that controls the amount of regularization in the minimization. The least-squares estimate of the weights using Tikhonov regularization is then

$$\beta_x = \left(\Phi_x^T \Phi_x + \lambda I \right)^{-1} \Phi_x^T s_x^{man},$$

where I is the identity matrix, a result which is analogous to ridge regression.

The second form of regularization was a spatial-neighborhood sampling scheme, which used local neighborhood samples as additional training samples in the least-squares estimation. Let θ_x^r denote the local neighborhood of radius r centered at spatial location x. We used the $\eta = (2r+1)^3$ local neighborhood segmentation samples, $S(\theta_x^r)$ as multiple training samples to estimate the optimal weights at x. The design matrix at each x, $\Phi_{\theta_x^r}$, then becomes an $\eta N \times M$ matrix, and $\mathbf{s}_{\theta_x^r}{}^{man}$ becomes an ηN-length column vector populated with the neighborhood training samples. The weights determined in this fashion, $\beta_{\theta_x^r}$, incorporate an added spatial regularity due to the overlapping neighborhoods at each spatial location.

Dynamic weight selection

Generalizing to additional images with a purely supervised approach may be questionable, since the atlas weights are completely dependent on the training images, i.e. it is a "static selection" method. An alternative approach would use dynamic information, for example, information obtained from the target-atlas pairing, to provide dynamic weights for fusion. If we represent the dynamic weights for the $i = \star$ target as $\gamma_{\star,j}$, the weighted fusion can be written as:

$$S_\star^{Dyn}(x) = \sum_{j=1}^{M} \gamma_{\star,j}(x) S_{\star,j}^{auto}(x).$$

Here, S^{Dyn} represents the weighted-fusion segmentation using dynamic information, γ, for weighting.

One choice for the dynamic weights, γ, are local estimates of registration accuracy, as was done recently (Isgum et al., 2009; Artaechevarria et al., 2009; van Rikxoort et al., 2010; Wu et al., 2007), made on the assumption that registration accuracy directly relates to label propagation accuracy. An explicit formulation of our registration accuracy estimate is given in Registration and accuracy estimates section and is most similar to that of (Artaechevarria et al., 2009) for comparative purposes. The estimates commonly used in practice, such as intensity differences post-registration, which we use, or normalized mutual information, are heuristic measures and not directly related to registration accuracy, thus direct application to dynamic weighting can be sub-optimal. However, if one combines a supervised approach with dynamic weighting, the heuristic aspect can be accounted for by learning how the weights are best applied for a training set.

Supervised and dynamic weight selection (SuperDyn)

One way to incorporate dynamic information into a supervised learning model is to use an effective atlas weight composed of both static and dynamic components, $\beta_j(x) \cdot \gamma_{\star,j}(x)$, where β is the static component, dependent on the training set, and γ is the dynamic component, dependent on $I_\star$. With this combined model, the linear system we obtain is similar,

$$e_i(x) : S_i^{man}(x) = \beta_0(x) + \sum_{j=1}^{M} \beta_j(x) \gamma_{i,j}(x) S_{i,j}^{auto}(x); i = 1 \ldots N.$$

The weighted fusion then proceeds similarly at each voxel location, x, for target image $i = \star$:

$$S_\star^{SuperDyn}(x) = \sum_{j=0}^{M} \beta_j(x) \gamma_{\star,j}(x) S_{\star,j}^{auto}(x); S_{\star,0}^{auto}(x) = 1.$$

This combined supervised and dynamic (SuperDyn) approach possesses the benefits of both approaches, performing optimally well for subjects similar to the training set, with the ability to dynamically modify its weights for new images.

Jackknife approach

Now to apply this supervised approach using one set of labeled images, or atlases, where there is no distinction between a training subject and an atlas subject, we apply a jackknife approach to learn weights for subsets of the group using the remainder as training subjects.

We partition a set of labeled N subjects into K subsets, D_k, $k=1\cdots K$, where each subset contains approximately M subjects. Then for the k-th partition, D_k, the subjects in the remaining partitions, $D_{\bar{k}}$, are used as "training" subjects to train weights for each "atlas" subject in D_k, which we denote as ${\beta_{j\in D_k}}^k$ along with the bias parameter β_0^k.

We would like to combine the weights from all partitions. First, suppose we use the k-th partition and its weights to segment an additional image, $I_\star$. The fused segmentation based on the k-th partition would be

$$S_\star^{Super,k}(x) = \sum_{\substack{j=0 \\ j\in\mathcal{D}^k}} \beta_j^k(x) S_{\star j}^{auto}(x).$$

To use all the partitions, we can apply equal-weighted fusion on the K segmentations from each partition:

$$\begin{aligned}\hat{S}_\star^{Super}(x) &= \frac{1}{K}\sum_{k=1}^{K} S_\star^{Super,k}(x),\\ &= \frac{1}{K}\sum_{k=1}^{K}\left[\sum_{\substack{j=0 \\ j\in\mathcal{D}^k}} \beta_j^k(x) S_{\star j}^{auto}(x)\right],\\ &= \sum_{\substack{j=0 \\ j\in\mathcal{D}^k}}\left[\left(\frac{1}{K}\sum_{k=1}^{K}\beta_j^k(x)\right) S_{\star j}^{auto}(x)\right].\end{aligned}$$

Thus we see that the overall weights, $\hat{\beta}_{j\in\mathcal{D}}$, can be expressed as an average of the weights from each partition:

$$\hat{\beta}_{j\in\mathcal{D}}(x) = \frac{1}{K}\sum_{k=1}^{K}\beta_j^k(x).$$

Note that the choice of K dictates the number of atlases in each partition and accordingly the number of training subjects available for to partition. Setting $K=N$ would train each atlas subject individually, using the remaining N-1 subjects for training, akin to (Sdika, 2010), whereas setting K lower would allow for ensemble learning at the expense of fewer training subjects. By learning the atlas weights concurrently we effectively use the team performance to choose the weights, instead of individual atlas performance, which has been shown to be preferred in studies of classifier selection (Ruta and Gabrys, 2005).

Label propagation and pre-processing

We performed label propagation segmentation using FS + LDDMM (Khan et al., 2008), which performs large deformation diffeomorphic metric mapping (LDDMM, (Beg et al., 2005)) registration on sub-region MRI volumes, using Freesurfer (FS) subcortical segmentations to aid in sub-region selection, affine registration, intensity normalization, and LDDMM registration. Registration accuracy estimates were then generated using these mappings, and these were all spatially normalized to the template space using a local affine transformation. A bounding box enclosing each hippocampus was used to define the region in these transformed images where weights are learned.

Freesurfer subcortical labeling

The freely-available Freesurfer volume-based pipeline (version 4.5.0) was used to generate initial subcortical labels for each image. Briefly, this processing includes removal of non-brain tissue using a hybrid watershed/surface deformation procedure (Ségonne et al., 2004), automated Talairach transformation, and segmentation of the subcortical white matter and deep gray matter volumetric structures (including hippocampus, amygdala, caudate, putamen, ventricles) (Fischl et al., 2002, 2004).

FS + LDDMM pre-processing

We first note that each hemisphere, including its associated labels, was processed separately throughout the FS + LDDMM processing, that is, non-rigid registration was performed for each hemisphere separately. First, the hippocampus and amygdala FS labels were extracted and used in an intensity-based affine registration of the target image to the atlas image, thus ensuring that the local volume of interest is well-aligned for subsequent non-linear registration. A bounding box, predefined in the atlas space using the extents of the atlas hippocampus and amygdala FS labels plus 12 voxel padding, was used to generate a sub-volumes region-of-interest (ROI). Intensity normalization of the atlas and target subvolume MRI images involved 1) percentile-based rescaling to eliminate outlying voxel intensities (0.5% to 99.5% percentiles), and 2) linear piecewise intensity normalization of the target to the atlas by globally matching the median cerebrospinal fluid (CSF) and white matter (WM) intensities. Median tissue intensities were found using the eroded FS labels of the relevant CSF and WM structures.

Diffeomorphic registration

Diffeomorphic registration methods are desirable in many atlas-based segmentation applications because of their inherent smoothness and ability to model large and fine-scale displacements. We used LDDMM (Beg et al., 2005), which generates diffeomorphic transformations by minimizing the following energy functional:

$$E_{i,j}(v) = \int_0^1 \|v_t\|_2^V \,\mathrm{d}t + \lambda\left\|I_j\left(\phi_{i,j}^{-1}\right) - I_i\right\|_{L^2}^2 \quad (1)$$

where v_t is a time-dependent vector field that is integrated to find the mapping, $\phi_{i,j}$, and I_j and I_i are the atlas and target images respectively. The mapping, $\phi:\Omega\to\Omega$, is smooth and has a smooth inverse, thus anatomy is mapped consistently, without fusions or tears, while preserving smoothness of anatomical features. We denote LDDMM registration as a function, $\mathrm{LDDMM}:(I_i,I_j)\to\phi_{i,j}$, which takes two images and outputs the diffeomorphic map registering the two image volumes.

LDDMM was performed in a multi-stage fashion, each stage using different image pairs and with each subsequent stage initialized with the velocity vector field, v_t, of the previous stage. In the first stage, the FS labels of the hippocampus, amygdala, and lateral ventricles were used, in the second stage, the Gaussian-smoothed ($\sigma=5$) MRI images were used, and in the final stage the non-smoothed MRI images were used. The velocity vector fields were discretized into 5 timesteps.

Finally, the atlas segmentations for each hemisphere were propagated to the target by applying the LDDMM and affine transformation, using linear interpolation to maintain precision when resampling the segmentations.

Registration accuracy estimates

We used local estimates of registration accuracy to inform the dynamic component of weight selection, specifically the squared intensity difference of the images after registration, as previous work has shown it to improve accuracy (Isgum et al., 2009; Artaechevarria et al., 2009; van Rikxoort et al., 2010). The registration accuracy estimate at each spatial location, x, in the i-th image, when the j-th atlas is registered to it, is computed as:

$$\gamma_{i,j}(x) = exp\left[-\frac{1}{\rho^2}\left(I_j\left(\phi_{i,j}^{-1}(x)\right) - I_i(x)\right)^2\right].$$

Here we are comparing the intensity of the target image $I_i(x)$ at each location, x, to that of the deformed atlas image, $I_j(\phi_{i,j}^{-1})$. The positive parameter ρ determines what magnitude intensity differences between the images will reduce γ from its maximum value of 1. For images with intensities ranging from 0 to 255, we set this to approximately 5% of the dynamic range, or $\rho = 15$.

Spatial normalization

In the derivation for the spatially-local weight selection, we assumed the images were sufficiently spatially normalized such that corresponding voxel locations among subject images should imply corresponding anatomy; we describe how this was achieved here. A template subject, $\{I_{j_0}, S_{j_0}\}$ was chosen from the set of atlas subjects, and each training and test image was linearly registered to it. Using a whole brain linear registration may not sufficiently align the substructures of interest, especially those in the medial-temporal lobe. We thus instead find a 12-parameter affine transformation, $\hat{T}_{i,j_0}$, for each structure of interest (e.g. the left hippocampus), to align the equally-weighted segmentation, S_i^{Eql}, to the template manual segmentation, $S_{j_0}{}^{man}$:

$$\hat{T}_{i,j_0} = \underset{T_{i,j_0}}{\operatorname{argmin}} \left\| S_{j_0}^{man} - T_{i,j_0}\left(S_i^{Eql}\right) \right\|,$$

using a standard gradient descent scheme initialized with a center-of-mass translation. These transformations were then applied to the training manual segmentations, automated segmentations, and registration accuracy estimates to bring all the data required for weight selection in a common reference frame. Because we are only performing a low-dimensional affine registration, we do not expect the choice of template to have a significant effect on the resulting alignment. The spatial normalization and pre-processing steps are summarized in Fig. 1.

STAPLE

We compared our segmentation fusion methods with the simultaneous truth and performance level estimation (STAPLE) algorithm (Warfield et al., 2004), an iterative EM algorithm that computes a probabilistic estimate of the true segmentation and a measure of the performance level represented by each segmentation. We used the Insight Segmentation and Registration Toolkit (ITK, http://www.itk.org) filter implementation of STAPLE with the default parameters, which also automatically calculates the prior positive classification probability as the average fraction of the image volume filled by the target object in each input segmentation. We applied STAPLE by using it to fuse the multiple propagated atlases. One limitation of this algorithm is that it can only accept discretely-labeled input images, thus input segmentations had to be thresholded prior to input into STAPLE, which is not the case for the other methods.

Global segmentation performance assessment

We assessed the performance of our proposed segmentation method and comparable methods by evaluating how close the resulting segmentations are to the corresponding manual segmentations. The most commonly used metrics are spatially-global, and measure the overlap between the segmentations, the difference in volumes, or the aggregate distances between the surface contours.

For volumetric overlap, we used the Dice overlap, also referred to as the mean overlap or the similarity index, which is computed between two binary segmentations as:

$$Dice\left(S^{auto}, S^{man}\right) = 2\frac{V\left(S^{auto} \cap S^{man}\right)}{V(S^{auto}) + V(S^{man})},$$

where $V(\cdot)$ is the volume of the segmentation, or the number of voxels. Segmentations were thresholded at zero to generate the binary segmentations required for this metric. A Dice overlap of 1 indicates complete volumetric overlap, and 0 indicates no overlap at all.

For a surface-distance measure, we used the mean surface distance, $SD(M_{auto}, M_{man})$, computed between two surface meshes as:

$$SD(M_{auto}, M_{man}) = \frac{1}{2}[sd(M_{auto}, M_{man}) + sd(M_{man}, M_{auto})], \quad (2)$$

$$sd(M_{auto}, M_{man}) = \frac{1}{N_{M_{auto}}} \sum_{a \in M_{auto}} \min_{m \in M_{man}} d(a, m). \quad (3)$$

Here, the segmentations are represented as surfaces, generated by a mid-intensity iso-surface of the segmentations, and $d(a,m)$ is the Euclidean distance between nodes a and m in surfaces M_{auto} and M_{man} respectively. The mean surface distance measures the mean deviation between M_{auto} and M_{man} over all surface nodes. Note that this measure does not require corresponding surface nodes, as closest point distances are used.

Manual volumetry is still widely used to examine structural change in MRI. To compare it to automated volumetry we generated Bland–Altman plots, examined the volume distributions for each

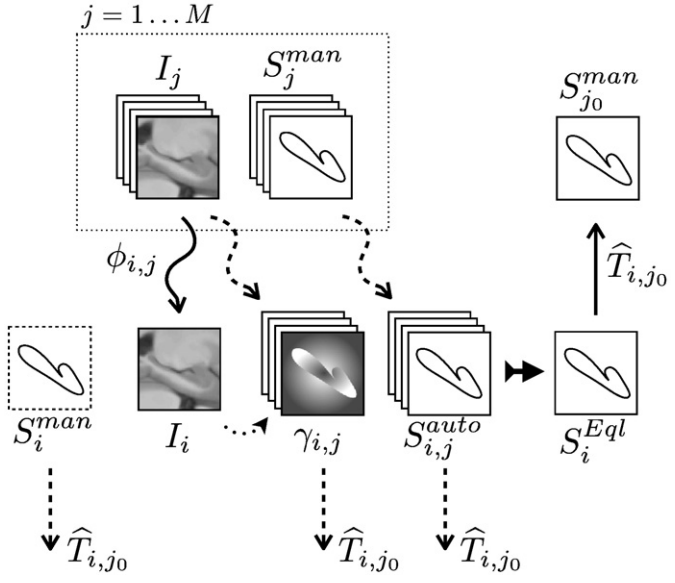


Fig. 1. Diagram showing how the atlas-propagation, registration accuracy estimates, and spatial normalization are carried out to bring a subject's data into the template space, where weight learning is carried out. The solid and dashed arrows represent registration and transformation respectively, with wavy lines for non-rigid and straight lines for affine.

method, computed Pearson correlations to the manual volumes, and computed an absolute volume error metric,

$$E_{vol} = \frac{|V(S^{auto}) - V(S^{man})|}{V(S^{man})} \times 100\%,$$

indicating the absolute difference in manual and automated volumes as a proportion of the manual volume.

Local surface-based segmentation performance assessment (SurfSPA)

The volumetric overlap and surface distance metrics provide information regarding how close or far the automated segmentation is to the manual at a global scale, but they do not provide information at a spatially-local scale. A spatially-local assessment of segmentation accuracy becomes even more relevant when the segmentations are to be used in subsequent shape analysis to detect subtle structural changes. Model-based methods have been used in the past (Yushkevich et al., 2006; Styner et al., 2004) to quantify and visualize the fit of statistical shape models of the hippocampus using boundary displacements. Here, we propose to use a similar method for generating local segmentation accuracy maps, which uses non-rigid image registration to generate corresponding surface models for each subject and each segmentation, then uses displacements between corresponding nodes to determine the bias and precision of a given segmentation method.

Hyper-template generation

We first generate a hyper-template to base our surface model on, representing the average shape of all the segmentations in our dataset. Using the approach as described in (Joshi et al., 2004; Khan and Beg, 2009), for each structure we generate the unbiased average segmentation of S_i^{man}. We use segmentations that have been linearly aligned to the chosen single subject template, S_{j_0}, and iteratively alternate between the following two steps at each iteration $k = 1 \ldots K$:

$$\text{Step 1}: \phi_{S_i,\ \bar{S}}(k) = LDDMM\left(S_i^{man},\ \bar{S}^{(k-1)}\right),$$

$$\text{Step 2}: \bar{S}^{(k+1)} = \frac{1}{N}\sum_{i=1}^{N} S_i^{man} \circ \phi^{-1}_{S_i,\ \bar{S}}(k),$$

with $\bar{S}^{(0)} = \frac{1}{N}\sum_{i=1}^{N} S_i^{man}$. The final average segmentation, $\bar{S}$, was then used to generate a surface mesh, $\bar{M}$, using a standard marching-cubes iso-surface algorithm.

Surface-injection

Once the surface model has been generated, we performed non-rigid registration (LDDMM) from the hyper-template segmentation image, $\bar{S}$, to each segmentation image, S_i^{method}, to obtain the mapping

$$\phi_{\bar{S},S_i^{method}} = LDDMM\left(\bar{S}, S_i^{method}\right).$$

The injected-surface was then computed as:

$$\widehat{M}_i^{method} = \phi^{-1}_{\bar{S},S_i^{method}}\left(\bar{M}\right).$$

Each $\widehat{M}$ now has the same set of corresponding nodes, obtained from the hyper-template mesh, $\bar{M}$, and thus all segmentations from different methods or subjects can be compared at a node-wise level. One benefit of this surface injection technique is that it can deal with many types of topological defects that can be present in manual or automated segmentations by enforcing a smoothness in the deformation that ignores holes or handles, as was done in(Qiu and Miller, 2008). However, this surface-based approach may not be appropriate for all types of hippocampal shapes depending on how much atrophy or segmentation defects are present.

Local segmentation accuracy statistics

Given these corresponding meshes for each subject and each type of segmentation, we would like to see where the automated segmentations under-segment or over-segment the manual segmentation, or equivalently, where the automated surface lies inside or outside the manual surface. To accomplish this, at each node, $a \in \widehat{M}^{auto}$, $m \in \widehat{M}^{man}$, we find the dot-product of the displacement from manual to auto, $\vec{d}_{m,a}$ and the surface normal of the manual, $\vec{n}_m$,

$$d^{norm}(m,a) = \vec{d}_{m,a} \cdot \vec{n}_m.$$

The normal distance, d^{norm}, is negative or positive when the automated surface is respectively inward or outward relative to the manual, effectively an indication of segmentation accuracy. Fig. 2 shows a simplified schematic of how these normal distances are computed between at two points on a pair of curves.

Statistical analysis of these distances can then be performed across the test subjects, including simple descriptive statistics such as the mean segmentation accuracy. In this case, a mean close to zero implies the automated segmentation is, on average, close to the manual segmentation, or that there is no bias in the boundary. However a more noteworthy characteristic is consistency, since volumetry or morphometry are unlikely to be adversely affected if the automated segmentations consistently overestimate or underestimate certain regions. The variance of the segmentation accuracy effectively measures this property, with lower variances implying greater precision and less segmentation noise that may confound or reduce the sensitivity of further shape analysis. We can also examine the distribution of these measures over the whole surface to better quantify differences between segmentation methods and their locations on the structure.

Materials

The first dataset comprised of a subset of 69 subjects from the PATH Through Life project, a longitudinal study of cognitive decline in healthy middle-aged subjects (Cherbuin et al., 2009). These subjects ranged in age from 44 to 49, and consisted of 39 males and 30 females. MRI scans were acquired for each subject on a 1.5T Gyroscan scanner (ACS-NT, Philips Medical Systems). T1-weighted images were acquired in the coronal plane using a Fast Field Echo (FFE) sequence (TR = 8.84 ms, TE = 3.55 ms, flip angle = 8°, matrix = 256 × 256, slices = 160, FOV = 256 × 256 mm, slice thickness = 1.5 mm).

To evaluate performance of our methods on an older group of subjects, we used a supplementary set of 37 elderly subjects from part

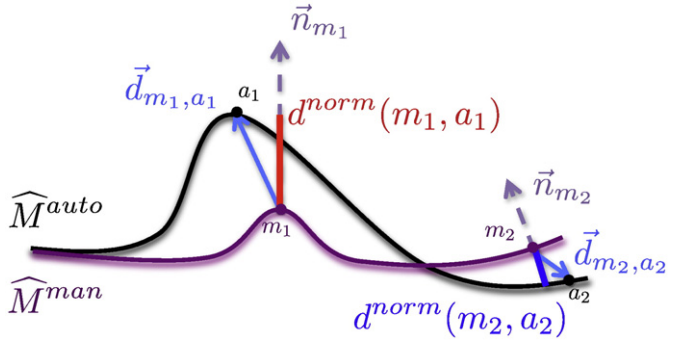


Fig. 2. Schematic showing how the normal distances, $d^{norm}(m,a)$ are computed between a pair of corresponding surfaces, $\widehat{M}^{auto}$ and $\widehat{M}^{man}$. The normal distance is the scalar defined as the dot-product between the manual surface normal, $\vec{n}_m$, and the manual to automated displacement vector, $\vec{d}_{m,a}$, which is thus positive or negative when the automated surface is outward or inward relative to the manual.

of the Memory and Aging Study (MAS) program, which comprises community residents aged between 70 and 90 recruited randomly through electoral roll from two electorates of East Sydney, Australia. The subjects ranged in age from 72 to 84, consisted of 17 males and 20 females, with 19 diagnosed with mild cognitive impairment (MCI) and 18 healthy controls. Subjects were scanned using a Philips 3T Intera Quasar scanner (Philips Medical Systems, Best, The Netherlands), and T1-weighted images were acquired (TR = 6.39 ms, TE = 2.9, ms, flip angle = 8°, matrix size = 256 × 256, slices = 190, FOV = 256 × 256 mm, slice thickness = 1 mm).

Left and right hippocampi were manually outlined on coronal slices using Analyze 5.0 (Brain Imaging Resource, Mayo Clinic, Rochester, MI, USA) according to the protocol described by (Watson et al., 1997). The hippocampal tail was manually traced according to the protocol detailed in (Maller et al., 2006). The manual rater for the study was NC and the scans of ten individuals were re-traced to compute a intra-class correlation (ICC) measures, which were 0.990 for the left and 0.997 for the right hippocampus (Cherbuin et al., 2009). An inter-class correlation measure was computed for another sample from the same larger study using the same manual rater and demonstrated very high inter-rater reliability (Maller et al., 2006).

The protocol does not differ significantly from that used by Freesurfer in definition, however in practice the subicular/entorhinal and parahippocampal boundary is assessed differently (Cherbuin et al., 2009).

Experiments

We performed segmentation of the subjects using cross-validation on all subjects on each dataset, (PATH, MAS) using the following segmentation methods:

- *Eql:* Multi-atlas FS + LDDMM with equally weighted fusion (majority voting)
- *STAPLE:* Multi-atlas FS + LDDMM with STAPLE EM-based fusion
- *Dyn:* Multi-atlas FS + LDDMM with dynamic (registration accuracy) weighted fusion
- *Super:* Multi-atlas FS + LDDMM with jackknife supervised-learning weighted fusion
- *SuperDyn:* Multi-atlas FS + LDDMM with jackknife supervised-learning + dynamic (registration accuracy) weighted fusion

We evaluated the fusion methods using $N = 30$ training subjects for propagation with the remaining subjects for testing. We randomly chose the training and test sets using leave-N-out cross-validation and performed 10 trials to obtain 10 different evaluation sets. The performance scores for each method and evaluation subject were averaged over the multiple randomized trials. For the jackknife supervised-learning approaches we fixed $M = 10$ by choosing $K = ceil\frac{N}{M}$, so that 10 atlases and 20 training subjects were used in each jackknife partition. Both Tikhonov and spatial neighborhood regularization were used, with the Tikhonov regularization parameter, λ, found using κ-fold cross-validation ($\kappa = 6$) on a fixed training set of 30 subjects, choosing the value that resulted in highest mean hippocampus Dice overlap, $\lambda = 10^{-3}$, and the spatial neighborhood, r, fixed to 1-voxel.

For the PATH dataset, we also performed experiments testing the effect of training set size by varying $N = \{10, 20, 30, 40\}$, and testing the team performance of the supervised learning methods by varying $M = \{3, 5, 10, 20\}$, with $M < N$.

Results

Fig. 3 shows summary box-plots of the volumetric overlap and surface distances versus the manual tracings for each segmentation method on the 69 subject PATH dataset. The results show STAPLE performs less satisfactorily than Eql, which is consistent with published reports(Artaechevarria et al., 2009; Sabuncu et al., 2010).

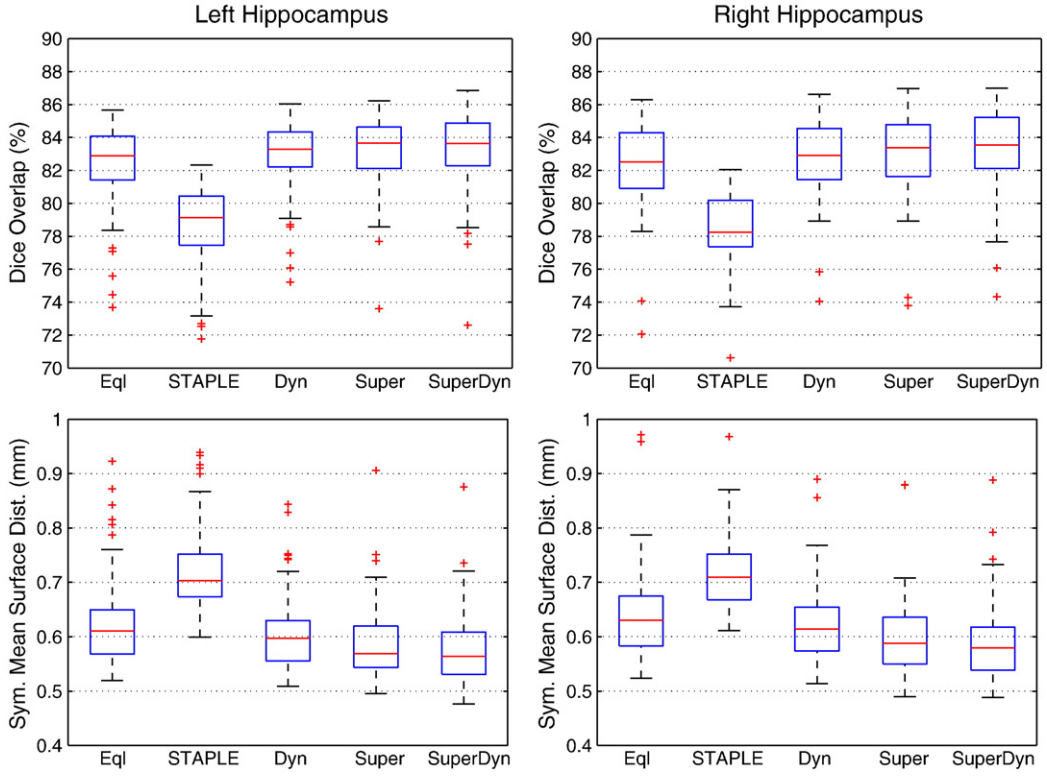


Fig. 3. Box-plots of performance metrics on the PATH dataset for segmentation methods, as compared against manual segmentations, for the left hippocampus (left column) and the right hippocampus (right column). The top row contains the Dice overlaps, higher values implying greater volumetric overlap with the manual segmentation, and the bottom row contains the mean surface distances, in mm, lower values implying a surface closer to the manual segmentation.

Table 1
Dice overlap metrics on the PATH dataset (mean ± standard deviation) for the various segmentation methods as compared with manual segmentations. Paired t-tests were performed to test for significantly different means versus the SuperDyn method, with t-statistics and p-values reported for each method.

Left hippocampus				Right hippocampus			
Method	Dice (%) $\mu \pm \sigma$	t-stat	p-val	Method	Dice (%) $\mu \pm \sigma$	t-stat	p-val
Eql	82.3 ± 2.6	5.8	<0.001	Eql	82.3 ± 2.6	6.2	<0.001
STAPLE	78.5 ± 2.5	23	<0.001	STAPLE	78.4 ± 2.2	29	<0.001
Dyn	82.9 ± 2.3	2.3	0.024	Dyn	82.8 ± 2.3	2.5	0.014
Super	83.1 ± 2.3	3.1	0.0027	Super	83.0 ± 2.4	3	0.0038
SuperDyn	83.3 ± 2.4	–	–	SuperDyn	83.2 ± 2.4	–	–

Table 2
Symmetric mean surface distance metrics on the PATH dataset (mean ± standard deviation) for the various segmentation methods as compared with manual segmentations. Paired t-tests were performed to test for significantly different means versus the SuperDyn method, with t-statistics and p-values reported for each method.

Left hippocampus				Right hippocampus			
Method	Surf. dist. (mm) $\mu \pm \sigma$	t-stat	p-val	Method	Surf. dist. (mm) $\mu \pm \sigma$	t-stat	p-val
Eql	0.63 ± 0.09	8.9	<0.001	Eql	0.64 ± 0.08	11	<0.001
STAPLE	0.72 ± 0.08	21	<0.001	STAPLE	0.72 ± 0.07	22	<0.001
Dyn	0.61 ± 0.07	6.6	<0.001	Dyn	0.62 ± 0.07	7.5	<0.001
Super	0.58 ± 0.07	3.4	0.0011	Super	0.60 ± 0.07	3.6	<0.001
SuperDyn	0.58 ± 0.07	–	–	SuperDyn	0.59 ± 0.07	–	–

Using dynamic information for weighting (Dyn) does improve accuracy with similar improvements in the supervised-learning methods, with the SuperDyn segmentations closest to the manual segmentation. Tables 1 and 2 show the mean values and standard deviations for the Dice and surface distance metrics respectively for the PATH dataset, along with the results of statistical tests (paired t-tests) to test significantly different means versus SuperDyn. Although the marginal improvement using SuperDyn over equal weighting may seem minor, a similar improvement with locally-weighted fusion was reported recently in (Artaechevarria et al., 2009). Furthermore, the relative improvement via atlas weighting also depends on how well the registration performs; if the registration is less accurate such that the number of misclassified voxels in atlas-based segmentation is increased, then the potential improvement via atlas weighting or other fusion strategies would thus also be increased. This effect could be quantified by varying the registration parameters to obtain highly-regularized (and thus less accurate) registrations, however we leave this for future work as it is beyond the scope of this article.

The supervised fusion methods outperformed the compared segmentation methods, however, we would also like to determine how the accuracy is affected by the training set size and the number of atlases used in each jackknife subset. The results of the experiments on the PATH dataset evaluating the effect of training set size and jackknife atlas set size are shown in Fig. 4. As expected we see there is an increasing trend in accuracy as training set size increases. We also see that increasing the jackknife atlas set size, M, also improves accuracy, but if M becomes too large, the number of training samples N–M becomes too small for effective learning, and thus over-fitting can occur.

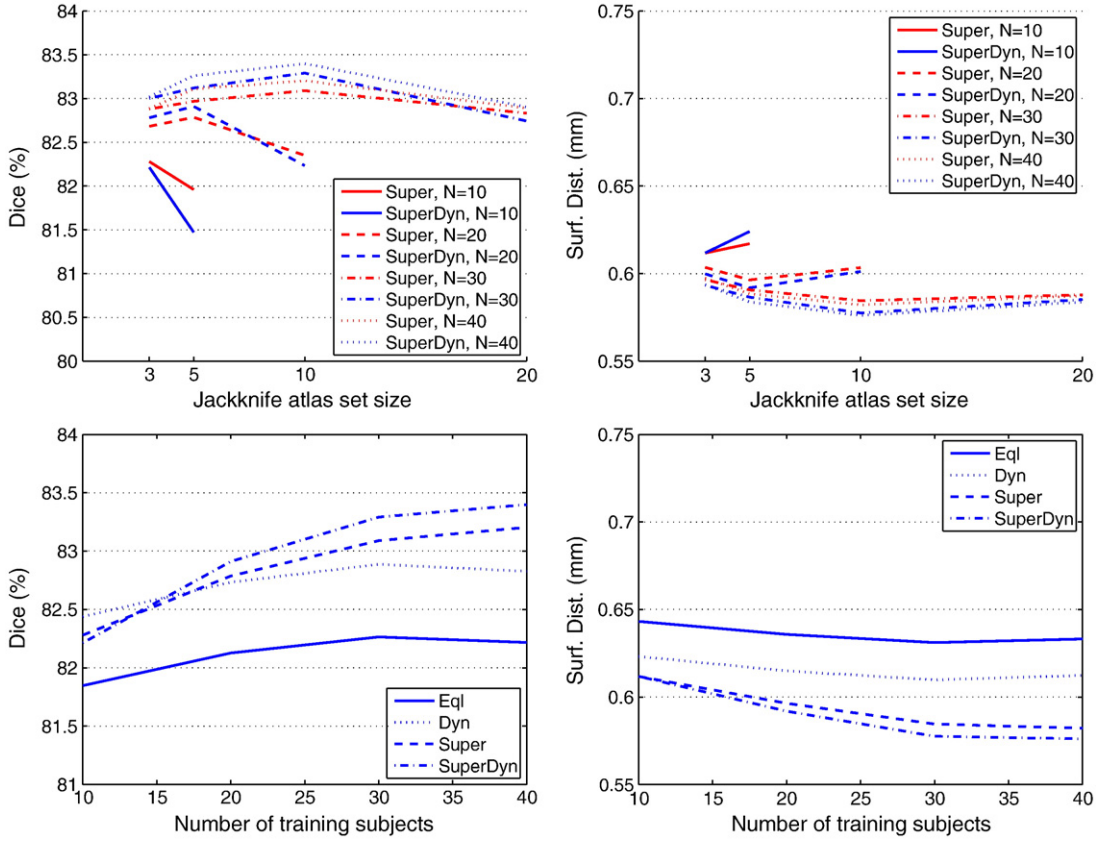


Fig. 4. Left hippocampus segmentation accuracy on the PATH dataset of the supervised-learning methods (Super, SuperDyn), evaluating the effect of the jackknife atlas subset size (M) and training set size (N) comparing to equally-weighted fusion (Eql) and dynamic fusion (Dyn). The N training subjects are used as propagation atlases. The top row shows the performance for the experiments for training size $N = \{10,20,30,40\}$, where the jackknife atlas size is varied,($M = \{3,5,10,20\}$, $M<N$). The bottom row compares Super and SuperDyn using the best performing M for each N with Eql and Dyn, which was $M = \{3,5,10,10\}$.

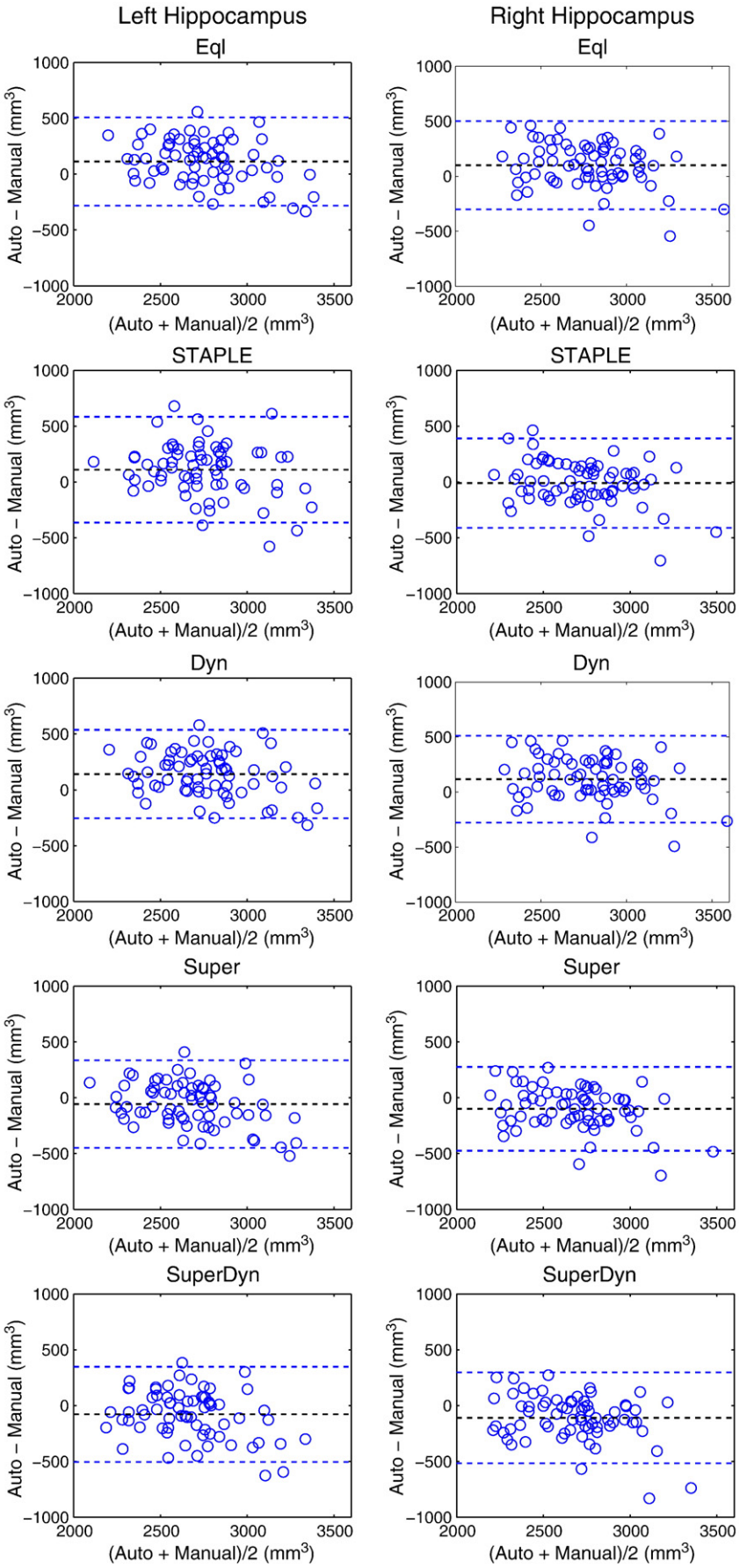


Fig. 5. Bland–Altman plots of the automated volume measurements showing graphically the agreement between automated and manual volumes. The black and blue dashed lines mark the mean difference and the mean difference $\pm 2SD$.

Table 3
Hippocampus volumes and absolute volume error (E_{vol}) measured on the PATH dataset shown for the various segmentation methods (mean ± standard deviation). Pearson correlations to the manual segmentation volumes are also shown for each method, with Super, Eql and Dyn showing the highest correlation with the manual volumes.

Left hippocampus				Right hippocampus			
Method	Vol. (mm^3) $\mu \pm \sigma$	E_{vol}(%) $\mu \pm \sigma$	Corr.	Method	Vol.(mm^3) $\mu \pm \sigma$	E_{vol}(%) $\mu \pm \sigma$	Corr.
Man	2700 ± 315	–	–	Man	2729 ± 313	–	–
Eql	2813 ± 249	7.2 ± 5.2	0.78	Eql	2829 ± 261	6.7 ± 5.4	0.77
STAPLE	2810 ± 262	7.9 ± 6.1	0.68	STAPLE	2719 ± 253	5.6 ± 4.6	0.77
Dyn	2841 ± 260	7.6 ± 5.6	0.78	Dyn	2846 ± 267	6.9 ± 5.5	0.78
Super	2643 ± 247	6.0 ± 3.9	0.78	Super	2630 ± 254	5.7 ± 4.5	0.80
SuperDyn	2621 ± 260	6.4 ± 4.6	0.74	SuperDyn	2619 ± 254	6.1 ± 5.0	0.76

Fig. 5 shows Bland–Altman plots of the volumes on the PATH dataset from each automated method with $N = 30$. From these we see that there is a general trend of all methods to underestimate the volumes of the larger hippocampi, which could be due to the effective spatial smoothing of the multi-atlas fusion segmentations related to the averaging over individual variations. To further understand how consistently the volumes are over- or underestimated, Pearson correlations with the manual volumes are shown in Table 3, along with the mean and standard deviation of volume and absolute volume error for each method. These show that the Super segmentations have volumes that are closest to the manual.

To evaluate performance on the elderly dataset (MAS) which exhibits more hippocampal atrophy, segmentations were computed with Dice overlaps and volumes summarized in Tables 4 and 5 respectively. We see that Dice overlaps are higher for this dataset than the PATH dataset, and that the supervised methods perform better than the others, as was the case on the PATH dataset. The overall increase in accuracy can be explained by the 3T versus 1.5T scanner difference, and thus greater signal-to-noise in the MAS scans.

To investigate segmentation accuracy at a spatially-local scale, we used SurfSPA to generate local accuracy statistics in a common surface model. Fig. 6 shows the mean and variance of the distance to the manual segmentations for each segmentation method, as colored on the hyper-template surface model. The mean indicates how much the automated segmentation under- or over-estimated the structure, and the variance indicates how much variability there is in this accuracy. Looking at the mean surface maps, we see the problematic regions of the hippocampus are mainly in the head region and at the boundary with the lateral ventricles, which could be because these regions may be more susceptible to partial-volume effects at the CSF boundary, especially for subjects with smaller ventricles where partial-volume effects can dominate.

As was seen in the global segmentation metrics, the mean distance maps show the STAPLE segmentation is farthest from the manual, Eql and Dyn are closer, and Super and SuperDyn are closer yet. The variance maps reveal that the segmentation methods are generally consistent for the majority of the hippocampal surface, with variance close to zero. For a more quantitative look at the performance assessment, we computed empirical cumulative distribution functions (eCDF) for the absolute mean and the variance of each segmentation method to summarize the spatial content. Fig. 7 shows these eCDF curves representing the cumulative distribution of mean and variance over the whole surface of the hippocampus. The supervised methods clearly have the lowest mean, however, the variance plots show reduced variability in all methods as compared to Eql, with improvements in the middle range for Super/SuperDyn/Dyn, and in the upper range for STAPLE. This can be explained by the problematic region of the head having reduced variability in STAPLE, and the lateral boundary with the ventricles having reduced variability in the other methods.

A visualization of a subset of the supervised weights from Super and SuperDyn is shown in Fig. 8. These weights were from one of the randomized cross-validation trials of the left hippocampus ($N = 30$, $M = 10$) with visualizations shown for 6 of the training subjects plus the bias term. We see that there are many regions in the SuperDyn weights where the atlas segmentations are heavily weighted towards the foreground or background. In the Super weights, the magnitude of the weights are not as high, but spatial similarities can be seen when these are compared. Note that the SuperDyn weights, β_i are further modulated by the dynamic weights $\gamma_{i,j}$ in fusion, which can explain the differences between the two, especially in the bias terms.

Table 4
Dice overlap metrics on the elderly MAS dataset (mean ± standard deviation) for the various segmentation methods as compared with manual segmentations. Paired t-tests were performed to test for significantly different means versus the SuperDyn method, with t-statistics and p-values reported for each method.

Left hippocampus				Right hippocampus			
Method	Dice (%) $\mu \pm \sigma$	t-stat	p-val	Method	Dice (%) $\mu \pm \sigma$	t-stat	p-val
Eql	83.8 ± 3.1	8.2	<.001	Eql	84.1 ± .3	7.5	
STAPLE	77.4 ± .7	11	<0.001	STAPLE	80.8 ± 4.1	9.5	<0.001
Dyn	84.2 ± 3.1	6.8	<0.001	Dyn	84.7 ± 3.1	5	<0.001
Super	84.9 ± 3.2	2.5	0.017	Super	85.2 ± 3.2	2	0.049
SuperDyn	85.1 ± 3.1	–	–	SuperDyn	85.3 ± 3.1	–	–

Discussion and conclusion

One limitation faced by supervised learning methods is the requirement of a sufficient amount of training data. If enough training data are not provided, over-fitting could occur, whereby the learned weights would be highly dependent on specific training individuals and not representative of the target cohort in general. Furthermore, as the model complexity is increased, that is, if more atlases are used in the jackknife approach, then more training data will be required. In Fig. 4 we see that increasing M, the number of atlases in each jackknife subset, improves the accuracy but if it is too high relative to N then the number of training samples will not be high enough and overfitting can occur. For training set sizes of $N = \{10, 20, 30, 40\}$, we found the optimal values to be $M = \{3, 5, 10, 10\}$, which means N should ideally be 3 to 4 times M, or equivalently there should be 2 to 3 training samples (N–M) for each atlas in the jackknife learning approach.

Another limitation of our proposed supervised method is that it works on single structures, that is, weights for each structure are determined individually. This choice was made mainly for computational efficiency, but is also related to the segmentation representation we have used, which is capable of only single label representation. We chose to use a continuous measure for segmentation representation instead of discrete labels to avoid loss of detail during interpolation. Alternatively one could use signed distance transform representations, such as LogOdds (Pohl et al., 2007) to represent the probabilistic segmentations. To incorporate multiple structures into our learning framework, we would also have to extend

Table 5
Hippocampus volumes and absolute volume error (Evol) measured on the elderly MAS dataset shown for the various segmentation methods (mean ± standard deviation). Pearson correlations to the manual segmentation volumes are also shown for each method, with Eql, Dyn, and SuperDyn showing the highest correlation with the manual volumes.

Left hippocampus				Right hippocampus			
Method	Vol. (mm^3) $\mu \pm \sigma$	E_{vol} (%) $\mu \pm \sigma$	Corr.	Method	Vol. (mm^3) $\mu \pm \sigma$	E_{vol} (%) $\mu \pm \sigma$	Corr.
Man	2432 ± 424	–	–	Man	2558 ± 523	–	–
Eql	2264 ± 320	7.4 ± 5.6	0.92	Eql	2354 ± 378	8.6 ± 5.5	0.96
STAPLE	2272 ± 347	11.4 ± 7.2	0.68	STAPLE	2524 ± 381	7.0 ± 9.9	0.94
Dyn	2268 ± 319	7.3 ± 5.7	0.92	Dyn	2367 ± 389	7.9 ± 5.3	0.95
Super	2344 ± 329	5.8 ± 5.2	0.91	Super	2447 ± 406	6.5 ± 4.8	0.95
SuperDyn	2340 ± 326	5.6 ± 5.3	0.92	SuperDyn	2442 ± 423	6.4 ± 4.4	0.95

our linear least squares solution to deal with multinomial variables representing the segmentation labels.

Several methods for intelligent atlas selection have been proposed recently for brain segmentation, many of them using dynamic information for selection (Artaechevarria et al., 2009; van Rikxoort et al., 2010; Wu et al., 2007; Aljabar et al., 2009), such as mean-squared intensity differences post-registration (as we have used), mutual information measures, or meta-data. In our experiments we have also found that using dynamic information alone (Dyn) does increase the Dice overlap versus equal weighted fusion (Eql) and when dynamic information is used with supervised learning (SuperDyn) all performance metrics show improvements; most importantly the variability (SurfSPA) is reduced. One possible explanation for this is that the arbitrary scaling of the dynamic weights in Dyn is sub-optimal, but when used in the supervised setting (SuperDyn), the optimal scaling factors become embedded in the learned weights,

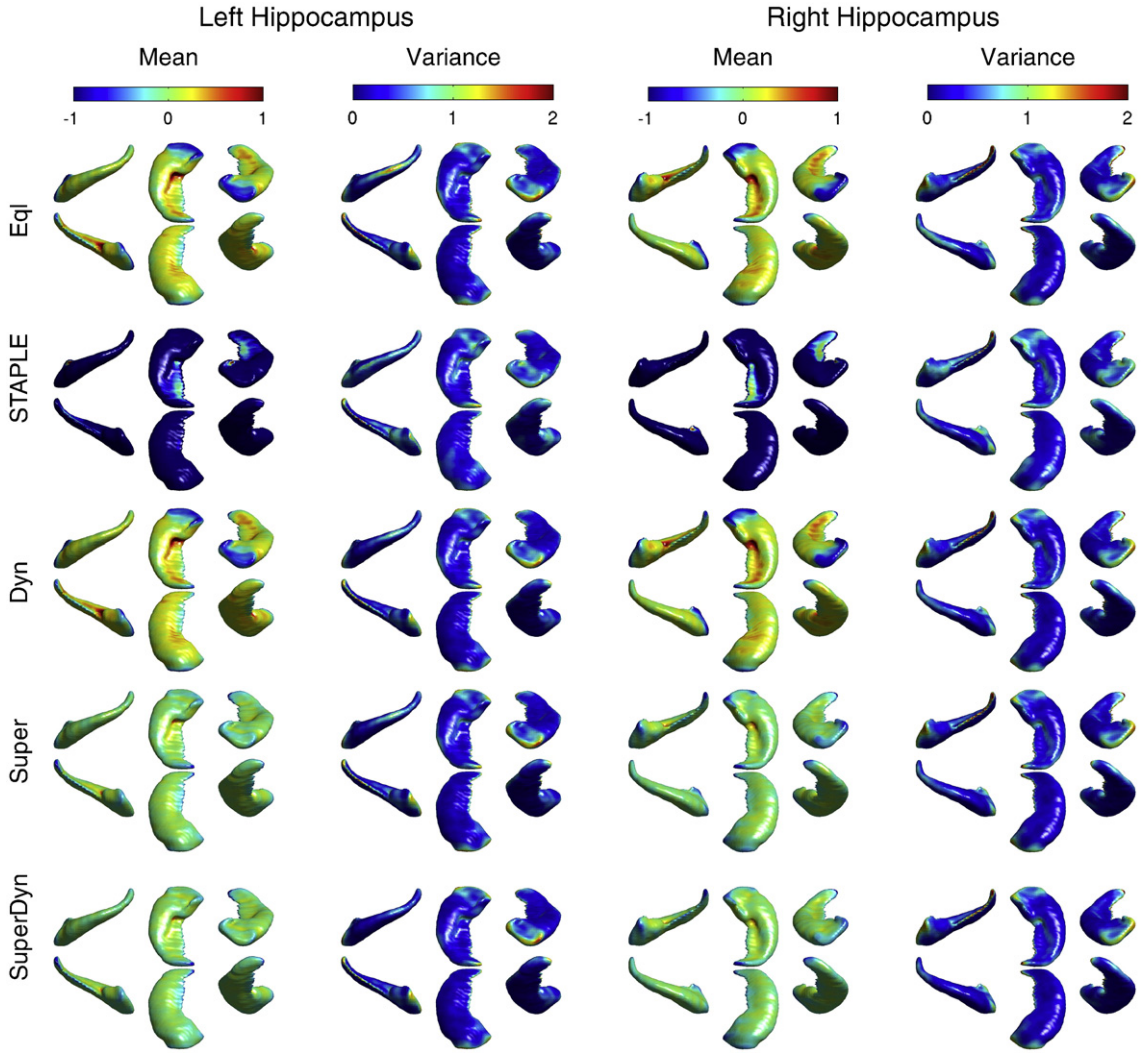


Fig. 6. Surface–surface displacement statistics on the PATH dataset showing the displacement from manual to automated segmentations, as averaged over the all the randomized cross-validation experiments ($N = 30$) displayed on an average hippocampus surface model. The visualizations in the mean columns indicate the mean inward/outward surface displacement to the manual (in mm), with negative values relating to an interior-placed automated surface, and positive values relating to an exterior-placed automated surface. The visualizations in the variance columns show the variance of these distance measures, indicating how consistently the automated segmentations under- or over-estimates the surface of the structure.

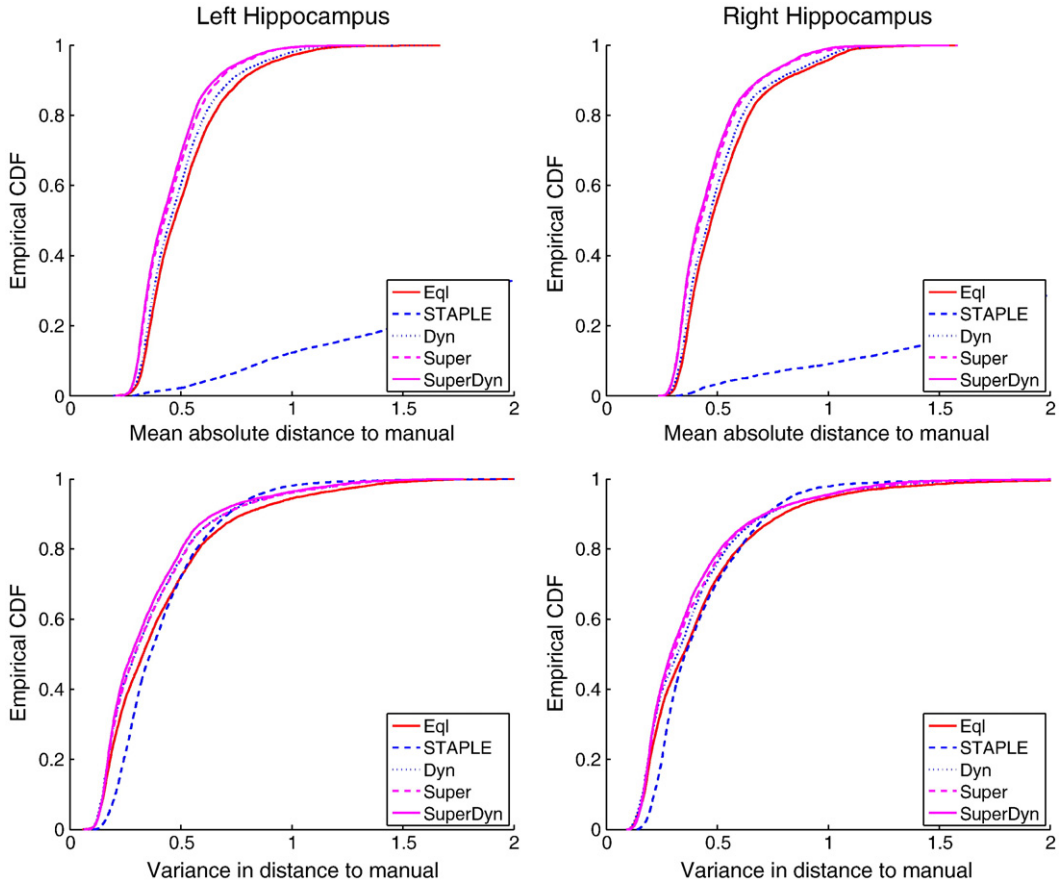


Fig. 7. Empirical cumulative distribution functions (CDFs) of surface displacement absolute mean (top row) and variance (bottom row), from SurfSPA metrics on the PATH dataset, shown in Fig. 6, summarize the spatial distribution of performance.

mitigating the heuristic nature of the dynamic weights. Future work could include evaluation of different types of dynamic information, both global and local, such as mutual information, shape similarity, or age, and how these measures perform alone or combined in a supervised setting.

Another difference in our locally-weighted methodology and recent work on this subject is regarding the spatial scale of the local weighting. Here, we have voxel-wise weights with a spatial neighborhood radius of 1 voxel. Other methods using dynamic fusion for brain segmentation use a radius of 5 (Artaechevarria et al., 2009) or larger bounding boxes encompassing the structures of interest (Wu et al., 2007; Aljabar et al., 2009). Considering the width of the hippocampus is close to 10 voxels itself, these methods are not effectively applying local weights at a sub-structure level of detail. Our dynamic fusion and SuperDyn methods are the only thus far to use local dynamic weights at a high-level of spatial resolution at the sub-structure level of detail, such that their use can improve the segmentation of single structures. Since other methods did not achieve good results overall using smaller spatial neighborhoods to segment multiple brain structures (Artaechevarria et al., 2009), it becomes evident that using adaptive neighborhood sizes could further improve performance of multiple structure local-weighted fusion. In this way, smaller neighborhoods could be used for structures such as the hippocampus, and larger neighborhoods could be used for structures with greater shape variability, such as those in the cerebral cortex.

Our surface-based segmentation performance assessment (SurfSPA) uses template surface injection to obtain corresponding surfaces. This technique was proposed by (Qiu and Miller, 2008) to perform shape de-noising of FS segmentations, a feature which is amenable to our segmentation performance assessment, since it allows for more stable comparison of segmentations that can suffer from manual tracing or binarization artifacts. One caveat with this method is that because each segmentation is effectively estimated

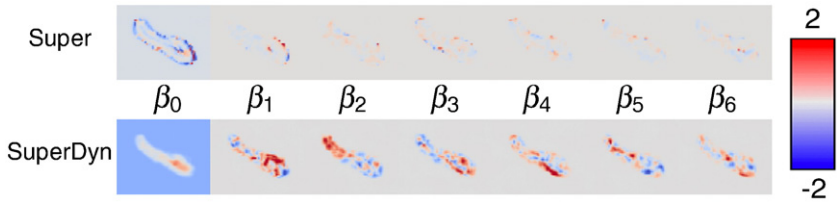


Fig. 8. Visualization of weights from the supervised learning methods, Super and SuperDyn. The representative weights were the selected from the first 6 training subjects in a randomly selected cross-validation trial of the left hippocampus ($N=30, M=10$) and the bias terms, β_0 from these experiments are also shown. Note that the SuperDyn weights are further modulated by the dynamic weights, $\gamma_{i,j}$, during optimal weight fusion.

with a diffeomorphic transformation of the hyper-template, there is the tendency that the injected segmentations will be more similar to each other than the original segmentations are, due to the inexactness of image matching. This would reduce overall surface-displacements expressing segmentation performance, and thus reduce the sensitivity of detecting differences in performance between segmentation methods; however in practice we are still able to quantify consistent differences between many of the segmentation methods we compared. To overcome this problem, a better surface matching could be obtained by extending a combined surface and volume registration method (SAVOR) (Gibson et al., 2009) to subcortical structures.

The SurfSPA methodology can also be used for deformation-based morphometry by analyzing the surface displacements relative to the hyper-template itself. In this way, one can visualize local group differences or correlations with socio-demographic or cognitive variables. Specifically, for future work we plan to explore how the spatially-local patterns of correlation with memory differ between the various segmentation methods (Manual, FS, Eql, and SuperDyn), as was done using hippocampal volumetry with manual and FS in (Cherbuin et al., 2009).

We tested our novel segmentation methods on 1.5T and 3T datasets containing 69 normative, middle-aged subjects, and 37 elderly MCI and normative subjects respectively, showing that the SuperDyn method significantly outperforms competing methods. The number of test subjects used is commonly accepted as being reasonable for demonstrating the validity of this method. In many practical scenarios, however, one would be seeking to identify either longitudinal hippocampal atrophy or comparing atrophy across different groups, such as healthy and demented. Thus, validation and application of our method to large datasets of patients with atrophied or abnormal hippocampi remains an outstanding issue which we plan to tackle in future studies. The PATH Through Life or MAS datasets could be used to this end, investigating changes in hippocampal atrophy between mild cognitive impairment (MCI) and healthy subjects.

We evaluated our segmentations on the two datasets separately, and thus separate sets of weights were learned for each. Considering generalizability, as with other supervised learning approaches the weights and atlases are dependent on the training set used. Thus although the atlases and their optimal weights established on one dataset could be used to segment another un-related test dataset, the resulting segmentation accuracy is likely be sub-optimal. The accuracy of segmentations in such an approach would be dependent on many factors, including how well the parameters of the test dataset such as scanner type, scan sequence, manual segmentation protocol, and patient age or disease state match the training dataset on which the atlases and weights were generated. The dynamic component of our proposed atlas weights would aid in generalizability, but overall, better results are likely be obtained when the test dataset is more similar to the atlases and the training dataset. In this respect, performing dynamic atlas ranking and selection, similar to that of (Aljabar et al., 2009), prior to atlas weight learning would also improve generalizability if a large and diverse atlas set was used.

In conclusion, we have introduced a supervised method of local atlas weighting in multi-atlas fusion, and also showed how dynamic information can be included to increase segmentation performance (SuperDyn). A jackknife learning approach was used to learn weights for a labeled training set. We validated our method using standard global segmentation metrics and a local surface-based method for segmentation performance analysis (SurfSPA), which was used to visualize accuracy maps of the hippocampus and quantify variability of the segmentation with respect to manual tracing. This analysis had shown that our novel Super and SuperDyn methodology improved overall accuracy and reduced local shape variability, though these methods require a larger number of manually-labeled training samples to avoid the effects of overfitting; as MRI databases become larger this trade-off becomes more favorable.

Acknowledgments

The authors are grateful to Anthony Jorm, Helen Christensen, Bryan Rodgers, Andrew Janke, National Capital Diagnostic Imaging group, Patricia Jacomb, Karen Maxwell, June Cullen, the Neuroimaging Group, NPI, Prince of Wales Hospital, and the PATH interviewers. We are also grateful to Jerome Maller and Chantal Meslin for the manual segmentations. We acknowledge Michael Smith Foundation for Health Research, Pacific Alzheimer Research Foundation and the National Health Medical Research Council, Australia for providing funding support for this work. A. R. Khan was supported by an NSERC CGS-D scholarship.

References

Aljabar, P., Heckemann, R.A., Hammers, A., Hajnal, J.V., Rueckert, D., 2009. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. Neuroimage 46 (3), 726–738.
Apostolova, L.G., Dinov, I.D., Dutton, R.A., Hayashi, K.M., Toga, A.W., Cummings, J.L., Thompson, P.M., 2006. 3D comparison of hippocampal atrophy in amnestic mild cognitive impairment and Alzheimer's disease. Brain 129 (Pt 11), 2867–2873.
Artaechevarria, X., Munoz-Barrutia, A., de Solorzano, C.O., 2009. Combination strategies in multi-atlas image segmentation: application to brain MR data. IEEE Trans. Med. Imaging 28 (8), 1266–1277.
Bajcsy, R., Lieberson, R., Reivich, M., 1983. A computerized system for the elastic matching of deformed radiographic images to idealized atlas images. J. Comput. Assist. Tomogr. 7 (4), 618–625.
Beg, M.F., Miller, M.I., Trouvé, A., Younes, L., 2005. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int. J. Comput. Vision 61 (2), 139–157.
Cherbuin, N., Anstey, K.J., Réglade-Meslin, C., Sachdev, P.S., 2009. In vivo hippocampal measurement and memory: a comparison of manual tracing and automated segmentation in a large community-based sample. PLoS ONE 4 (4), e5265.
Fischl, B., Salat, D., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33 (3), 341–355.
Fischl, B., Salat, D., van der Kouwe, A., Makris, N., Ségonne, F., Quinn, B., Dale, A., 2004. Sequence-independent segmentation of magnetic resonance images. Neuroimage 23, 69–84.
Gibson, E., Khan, A.R., Beg, M.F., 2009. A combined surface and volumetric registration (SAVOR) framework to study cortical biomarkers and volumetric imaging data. Medical image computing and computer-assisted intervention: MICCAI International Conference on Medical Image Computing and Computer-Assisted Intervention, 12(Pt 1), pp. 713–720.
Hammers, A., Heckemann, R., Koepp, M.J., Duncan, J.S., Hajnal, J.V., Rueckert, D., Aljabar, P., 2007. Automatic detection and quantification of hippocampal atrophy on MRI in temporal lobe epilepsy: a proof-of-principle study. Neuroimage 36 (1), 38–47.
Heckemann, R., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A., 2006. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. Neuroimage 33 (1), 115–126.
Isgum, I., Staring, M., Rutten, A., Prokop, M., Viergever, M.A., van Ginneken, B., 2009. Multi-atlas-based segmentation with local decision fusion–application to cardiac and aortic segmentation in CT scans. IEEE Trans. Med. Imaging 28 (7), 1000–1010.
Joshi, S., Davis, B., Jomier, M., Gerig, G., 2004. Unbiased diffeomorphic atlas construction for computational anatomy. Neuroimage 23 (Suppl 1), S151–S160.
Khan, A.R., Beg, M.F., 2009. Multi-structure whole brain registration and population average. IEEE Engineering in Medicine and Biology Society Conference—EMBC 2009, pp. 5797–5800.
Khan, A.R., Wang, L., Beg, M.F., 2008. Freesurfer-initiated fully-automated subcortical brain segmentation in MRI using large deformation diffeomorphic metric mapping. Neuroimage 41 (3), 735–746.
Leung, K.K., Barnes, J., Ridgway, G.R., Bartlett, J.W., Clarkson, M.J., Macdonald, K., Schuff, N., Fox, N.C., Ourselin, S., 2010. Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. Neuroimage 51 (4), 1345–1359.
Maller, J.J., Réglade-Meslin, C., Anstey, K.J., Sachdev, P., 2006. Sex and symmetry differences in hippocampal volumetrics: before and beyond the opening of the crus of the fornix. Hippocampus 16 (1), 80–90.
Pohl, K.M., Fisher, J., Bouix, S., Shenton, M., McCarley, R.W., Grimson, W.E.L., Kikinis, R., Wells, W.M., 2007. Using the logarithm of odds to define a vector space on probabilistic atlases. Med. Image Anal. 11 (5), 465–477.
Qiu, A., Miller, M.I., 2008. Multi-structure network shape analysis via normal surface momentum maps. Neuroimage 42 (4), 1430–1438.
Rohlfing, T., Brandt, R., Menzel Jr., R., C. R.M., 2004. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. Neuroimage 21, 1428–1442.
Ruta, D., Gabrys, B., 2005. Classifier selection for majority voting. Inf. fusion 6, 63–81.

Sabuncu, M., Yeo, B.T., Leemput, K.V., Fischl, B., Golland, P., 2010. A generative model for image segmentation based on label fusion. IEEE Trans. Med. Imaging 29 (10), 1714–1729.
Sdika, M., 2010. Combining atlas based segmentation and intensity classification with nearest neighbor transform and accuracy weighted vote. Med. Image Anal. 14 (2), 219–226.
Ségonne, F., Dale, A.M., Busa, E., Glessner, M., Salat, D., Hahn, H.K., Fischl, B., 2004. A hybrid approach to the skull stripping problem in MRI. Neuroimage 22 (3), 1060–1075.
Styner, M., Lieberman, J.A., Pantazis, D., Gerig, G., 2004. Boundary and medial shape analysis of the hippocampus in schizophrenia. Med. Image Anal. 8 (3), 197–203.
Tepest, R., Wang, L., Csernansky, J.G., Neubert, P., Heun, R., Scheef, L., Jessen, F., 2008. Hippocampal surface analysis in subjective memory impairment, mild cognitive impairment and Alzheimer's dementia. Dement. Geriatr. Cogn. Disord. 26 (4), 323–329.
Thompson, P.M., Hayashi, K.M., Zubicaray, G.I.D., Janke, A.L., Rose, S.E., Semple, J., Hong, M.S., Herman, D.H., Gravano, D., Doddrell, D.M., Toga, A.W., 2004. Mapping hippocampal and ventricular change in Alzheimer disease. Neuroimage 22 (4), 1754–1766.
van Rikxoort, E.M., Isgum, I., Arzhaeva, Y., Staring, M., Klein, S., Viergever, M.A., Pluim, J. P.W., van Ginneken, B., 2010. Adaptive local multi-atlas segmentation: application to the heart and the caudate nucleus. Med. Image Anal. 14 (1), 39–49.
Wang, L., Khan, A., Csernansky, J.G., Fischl, B., Miller, M.I., Morris, J.C., Beg, M.F., 2009. Fully-automated, multi-stage hippocampus mapping in very mild Alzheimer disease. Hippocampus 19 (6), 541–548.
Wang, L., Mamah, D., Harms, M.P., Karnik, M., Price, J.L., Gado, M.H., Thompson, P.A., Barch, D.M., Miller, M.I., Csernansky, J.G., 2008. Progressive deformation of deep brain nuclei and hippocampal-amygdala formation in schizophrenia. Biol. Psychiatry 64 (12), 1060–1068.
Wang, L., Miller, J.P., Gado, M.H., McKeel, D.W., Rothermich, M., Miller, M.I., Morris, J.C., Csernansky, J.G., 2006. Abnormalities of hippocampal surface structure in very mild dementia of the Alzheimer type. Neuroimage 30 (1), 52–60.
Wang, L., Swank, J.S., Glick, I.E., Gado, M.H., Miller, M.I., Morris, J.C., Csernansky, J.G., 2003. Changes in hippocampal volume and shape across time distinguish dementia of the Alzheimer type from healthy aging. Neuroimage 20 (2), 667–682.
Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging 23 (7), 903–921.
Watson, C., Jack, C.R., Cendes, F., 1997. Volumetric magnetic resonance imaging. Clinical applications and contributions to the understanding of temporal lobe epilepsy. Arch. Neurol. 54 (12), 1521–1531.
Wu, M., Rosano, C., Lopez-Garcia, P., Carter, C.S., Aizenstein, H.J., 2007. Optimum template selection for atlas-based segmentation. Neuroimage 34 (4), 1612–1618.
Yushkevich, P.A., Zhang, H., Gee, J.C., 2006. Continuous medial representation for anatomical structures. IEEE Trans. Med. Imaging 25 (12), 1547–1564.