Nonlocal regularization for active appearance model: Application to medial temporal lobe segmentation.

Shiyan Hu, Pierrick Coupé, Jens C Pruessner, D Louis Collins

Nonlocal Regularization for Active Appearance Model: Application to Medial Temporal Lobe Segmentation

Shiyan Hu[a,*], Pierrick Coupé[a,b], Jens C. Pruessner[a,c], and D. Louis Collins[a]

[a]McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, Canada

[b]LaBRI CNRS, UMR 5800 Université, Bordeaux, France

[c]McGill Centre for Studies in Aging, Faculty of Medicine, McGill University, Montreal, Canada

Abstract

The human medial temporal lobe is an important part of the limbic system, and its substructures play key roles in learning, memory, and neurodegeneration. The medial temporal lobe includes the hippocampus, amygdala, parahippocampal cortex, entorhinal cortex, and perirhinal cortex – structures that are complex in shape and have low between-structure intensity contrast, making them difficult to segment manually in magnetic resonance images.

This paper presents a new segmentation method that combines active appearance modeling and patch-based local refinement to automatically segment specific substructures of the medial temporal lobe including hippocampus, amygdala, parahippocampal cortex, and entorhinal/perirhinal cortex from MRI data. Appearance modeling, relying on eigen-decomposition to analyze statistical variations in image intensity and shape information in study population, is used to capture global shape characteristics of each structure of interest with a generative model. Patch-based local refinement, using nonlocal means to compare the image local intensity properties, is applied to locally refine the segmentation results along the structure borders to improve structure delimitation. In this manner, nonlocal regularization and global shape constraints could allow more accurate segmentations of structures.

Validation experiments against manually-defined labels demonstrate that this new segmentation method is computationally efficient, robust and accurate. In a leave-one-out validation on 54 normal

young adults, the method yielded a mean Dice κ of 0.87 for the hippocampus, 0.81 for the amygdala, 0.73 for the anterior parts of the parahippocampal gyrus (entorhinal and perirhinal cortex), and 0.73 for the posterior parahippocampal gyrus.

Keywords: segmentation, appearance modeling, nonlocal means, label fusion, medial temporal lobe structures.

1 Introduction

The medial temporal lobes (MTL) are an important part of the limbic system in humans and include the hippocampus (HC), amygdala (AG), and the parahippocampal gyrus with its substructures entorhinal cortex (ERC), perirhinal cortex (PRC), and parahippocampal cortex (PHC). These structures play important roles in learning, memory, and neurodegeneration (LeDoux, 1989; Barense et al., 2005; Baxter, 2009). The HC is the most frequently investigated component of the MTL because of its role in memory and contextualisation. The AG is strongly involved in emotional and social processing, in particular, fear and anxiety. The ERC is the main interface between the HC and the neocortex and plays an important role in the formation and optimization of spatial memories. The PRC is involved in visual perception and memory, and the PHC is involved in scene recognition and social context. Recently, MTL structures have received considerable attention due to their importance in neurological diseases and disorders (Cendes et al., 1993; Mori et al., 1997). For example, changes in hippocampal volume have been shown to be an important marker of the early stages of Alzheimer's disease and temporal lobe epilepsy (Jack Jr. et al. 1992, 1997; Fox et al., 1996; Duzel et al., 2005). Likewise, the parahippocampal gyrus, especially its substructures ERC and PRC, has been argued as an additional, possibly even superior, marker of neurodegeneration and dementia (Xu et al., 2000). Unfortunately, research evidence is sparse possibly due to the fact that manually segmenting substructures of the parahippocampal gyrus is complex and time consuming while automated techniques are not generally available. Because of the importance of these structures in neurodegeneration and the high time investment in performing manual segmentation, there is significant interest in developing accurate, robust, and reliable segmentation techniques to automatically extract these structures from magnetic resonance (MR) imaging for volume and shape analyses.

Manual segmentation is considered highly accurate and treated as the current gold standard. However, the technique is time consuming, requires anatomical expertise, and requires constant control of inter- and intra-rater variability. Hence, it is difficult to apply manual segmentation in studies involving

large numbers of subjects. To overcome the limitations of manual segmentation, many automatic segmentation techniques have been proposed, with most model-based segmentation techniques falling into the following three categories: deformable models (Shen et al., 2002), appearance-based models (Cootes et al., 1998; Klemencic et al., 2004; Patenaude et al., 2011), and atlas-based techniques (Fischl et al., 2002; Collins et al., 1995).

Deformable models use parametric or nonparametric methods to initialize contours or surfaces and then match them to the object boundaries (Ghanei et al., 1998; Shen et al., 2002). To avoid the mismatch between the model edge and the multiple edges in the image, Chupin et al. (2007, 2009) applied structure-specific morphometric rules based on prior knowledge of anatomical features derived from training data to segment the HC and AG. Cootes et al. (1995) incorporated a statistical parameterization into the deformable shape model. The statistical parameterization can be derived from training data but it often imposes global shape constraints, suggesting that the model can be deformed only in ways implied by the training data. This idea of incorporating the statistical shape model into the deformable shape templates resulted in active-shape models (Cootes et al. 1995) while the idea of building up both statistical shape and intensity models for shape and intensity led to active appearance models (Cootes et al., 1998; Duchesne et al., 2002). To avoid the manual identification of landmarks in training data (Cootes et al., 1998), we (Hu and Collins, 2007; Hu et al., 2011) integrated the level-set method into the appearance modeling and further integrated multi-contrast MR images into the segmentation to improve its robustness and accuracy. Recently, a similar method has been also proposed by Toth and Madabhushi (2012), where instead of multi-contrast MR images, multiple features derived from T2 MR images were integrated into the appearance modeling. Patenaude et al. (2011) placed appearance models within a Bayesian framework to better capture the probabilistic relationship between shape and intensity.

Atlas-based segmentation techniques have attracted attention for their high levels of accuracy (Fischl et al., 2002; Collins et al., 1995). Atlas-based techniques use a template (i.e., MR image with manual segmentation) as prior information to assist in providing automatic labels. Unlike the work of Collins et

al. (1995), where the manual labels in the template were propagated to the target image through an inverse spatial transformation, Fischl et al. (2002) developed another automatic label assigning technique based on the probabilistic information derived from templates. To avoid potential bias from using only one template, Heckemann et al. (2006) and Aljabar et al. (2009) proposed multi-atlas based methods with label fusion. They further improved segmentation efficiency by selecting several similar templates instead of all templates from a given library. Inspired by their work, Collins and Pruessner (2010) also incorporated label fusion into the multi-atlas warping and achieved very accurate results for automatic HC segmentation. Wang et al. (Wang et al., 2011) used the multi-atlas technique with error correction to yield the best-published results for HC segmentation with respect to the manual labels. Nevertheless, these techniques are sensitive to registration error and selection of the templates, as they generally assign the same weight to all templates in the segmentation procedure. More recently, Coupé et al. (2011) used a nonlocal means patch-based label fusion approach to weight the expert manual segmentation in a library of templates based on the intensity similarity between patches. Since its introduction, this method has been extended to the multi-scale framework (Eskildsen et al., 2012), the multi-point approach (Rousseau et al., 2011) and regression-based strategies (Wang et al., 2011b). Moreover, patch-based label fusion has been used in different contexts such as Alzheimer's disease detection (Coupé et al., 2012) and neurosurgical planning (Haegelen et al., 2012). A more detailed review of segmentation methods can be found in Table 1. Note that neither atlas-based nor patch-based methods explicitly incorporate global shape constraints into the segmentation.

To integrate global shape constraints into the segmentation and increase the local structure fitting, we developed a new fully automatic segmentation method that combined the active appearance model (AAM) and patch-based technique into a general two-stage segmentation framework. In the first-stage segmentation, the AAM is used to capture the statistical characteristics of shape and intensity information in the training data. Although the AAM does, in fact, take into account local geometry, its ability to recover fine details at structure borders is limited by the number of principal components used in the

model. Thus, there is often some “blurring” of the structure shape. This issue can be addressed by the nonlocal means patch-based technique, which is employed as a second-stage segmentation to locally refine the tentative segmentation results from the first-stage segmentation. To impose coarse global constraints and also to limit the number of voxels for local segmentation refinement, the second-stage local refinement is performed only on a structure boundary area identified by the first-stage segmentation. In this manner, global shape constraints and a local regularization can be well integrated and this integration can better enable accurate structure segmentations. In addition, the structure boundary area identified by the first-stage segmentation can also greatly reduce the search area for the structure border and greatly reduce the computational complexity in the second-stage segmentation as otherwise a large number of voxels requiring the local refinement would be needed and the computational complexity would be extremely high. Finally, it is important to note that while there have been a large number of publications describing different methods for HC and AG segmentation (see Table 1), we are aware of only one paper that addresses automatic segmentation of the PHC (Heckemann et al., 2006), and to the best of our knowledge, no methods have been published with validated automatic segmentation results on the entorhinal and perirhinal cortex (EPC), two substructures of the parahippocampal gyrus. This might have to do with the complex anatomical variation found between subjects in the area of the parahippocampal gyrus: essentially all of the substructures of this gyrus develop around the collateral sulcus, a highly variable fold in the MTL that can be interrupted, branched, or fused with neighbouring occipitotemporal and rhinal sulci (Pruessner et al., 2002). However, developing accurate automated segmentation techniques for this structure would allow for a more systematic investigation and assessment of the contribution of the substructures of the parahippocampal gyrus to memory and neurodegeneration.

The main contributions of this paper are as follows: 1) A two-stage segmentation to combine the appearance model and nonlocal means patch-based method to capture the global shape variation and to locally refine the segmentation by weighting the local signed distance functions; 2) Application of the

proposed two-stage segmentation method to segment all MTL structures. In comparison to the HC and AG, other MTL structures, like the PRC and PHC, have much greater anatomical variability and their segmentation has been considered difficult. Here, the two-stage segmentation method is shown to outperform the appearance modeling method alone or patch-based local refinement method alone in segmenting those structures; 3) Characterization of volume properties of all MTL structures in healthy young adults against hemisphere, age, and gender.

2 Method

2.1 Appearance Model-based Segmentation

Appearance model (AAM)-based segmentation applies the eigen-decomposition technique to gray-scale MR images and shape data to capture the statistical variations of the intensity and shape information of the training data. To minimize the differences in size, orientation, and position between subjects, both training and test MR images are linearly and then nonlinearly registered (Collins and Evans, 1997) to an unbiased nonlinear average template (ICBM152 2009c nonlinear asymmetric 1×1×1 mm template (Fonov et al., 2011)) within the volume of interest surrounding the MTL structures. Based on the eigenvectors derived from the training data, the final shape and gray level can be given by

$$\begin{aligned} \phi &= \bar{\phi} + \mathbf{P}_{\phi}\mathbf{Q}_{\phi}\overline{w}_s^{-1}\mathbf{c} \\ \mathbf{g}_{t_1} &= \bar{\mathbf{g}}_{t_1} + \mathbf{P}_{g,t_1}\mathbf{Q}_{g,t_1}\mathbf{c} \end{aligned} \tag{1}$$

where $\bar{\phi}$ is the mean of the signed distance functions of the training shapes, $\bar{\mathbf{g}}_{t_1}$ is the mean gray-level (intensity-level) of the normalized T1-weighted training images, $\mathbf{P}_{\phi}$ and $\mathbf{P}_{g,t_1}$ are the eigenvectors derived from the training shapes and training gray-scale images, respectively. $\mathbf{Q}_{\phi}$ and $\mathbf{Q}_{g,t_1}$ are the appearance eigenvectors as ways of jointly parameterizing a set of intermediate shape and intensity parameters. $\overline{w}_s^{-1}$ is a standard deviation balancing factor. To be specific, if we consider *M* training

images, the *i*-th shape and gray-scale training images can be represented as a linear combination of their corresponding eigenvectors, i.e.,

$$\phi_i = \bar{\phi} + \mathbf{P}_\phi \mathbf{b}_{\phi,i} \qquad (2)$$
$$\mathbf{g}_{t_1,i} = \bar{\mathbf{g}}_{t_1} + \mathbf{P}_{g,t_1} \mathbf{b}_{g,t_1,i}$$

where $\mathbf{b}_{\phi,i}$ is a vector of intermediate shape parameters while $\mathbf{b}_{g,t_1,i}$ is a vector of intermediate intensity parameters.

We can define two intermediate parameter matrices $\mathbf{B}_\phi$ =[$\mathbf{b}_{\phi,i}$, i=1,2,...,M] and $\mathbf{B}_{g,t_1}$ =[$\mathbf{b}_{g,t_1,i}$, i=1,2,...,M] and further group them into a super-matrix $\mathbf{B}$ in this form

$$\mathbf{B} = \begin{bmatrix} \mathbf{W}_s \mathbf{B}_\phi \\ \mathbf{B}_{g,t_1} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_\phi \\ \mathbf{Q}_{g,t_1} \end{bmatrix} \mathbf{C} \qquad (3)$$

where $\mathbf{W}_s$ =diag($w_{s,1}, w_{s,2}, \ldots, w_{s,M}$) is used to balance the sets of intermediate shape and intensity parameters in deriving the parameter eigenvector matrix $\mathbf{Q} = \left[\mathbf{Q}_\phi^T, \mathbf{Q}_{g,t_1}^T\right]^T$. Of the matrix $\mathbf{W}_s$, the *i*-th diagonal element $w_{s,i}$ is set to $\sigma_{g,t1,i} / \sigma_{s,i}$, a ratio between an intermediate intensity level parameter standard deviation $\sigma_{g,t1,i}$ and an intermediate shape parameter standard deviation $\sigma_{s,i}$. The standard deviation balancing factor $\overline{w}_s^{-1}$ in Eq.1 is set to the inverse of the mean of { $w_{s,i}$, *i*=1,2,…,M}. In Eq.3,

$\mathbf{C}$ is a matrix with each column being a vector of linear combination weight coefficients, also known as model parameters. For the *i*-th pair of shape and gray-scale training images, their corresponding model parameter vector is the *i*-th column vector of $\mathbf{C}$. Eq.1 is a generic representation with $\mathbf{c}$ as a set of model parameters. By adjusting the model parameters $\mathbf{c}$, different MR images and their corresponding shapes can be synthesized.

The segmentation for a given T1-weighted test MR image $\mathbf{I}_{t_1}$ is achieved by minimizing the gray-level difference between the test image and the image synthesized from Eq.1. The cost function in the least square measure can be written as:

$$\varepsilon = \sum_{j=1}^{N_p} (I_{t_1,j} - g_{t_1,j})^2 \tag{4}$$

where $I_{t_1,j}$ is the intensity of the *j*-th voxel of the T1-weighted test MR image $\mathbf{I}_{t_1}$, $g_{t_1,j}$ is the intensity of the *j*-th voxel of the synthesized T1-weighted image $\mathbf{g}_{t_1}$, expressed as a function of $\mathbf{c}$ given by Eq.1, and N_p is the total number of voxels in each image. The processing pipeline for the AAM-based segmentation is shown in Figure 1, and the segmentation method was described in detail in our previous work (Hu and Collins, 2007).

2.2 Nonlocal Means Patch-based Segmentation

To label a voxel in a test image, the nonlocal means patch-based segmentation procedure compares a small image patch from the test image to corresponding patches in a series of pre-labelled images in a template library. The label is obtained from a weighted average of the template labels. The method described here is the same as in Coupé et al., (2011). In particular, for a voxel x_i in the test image and a voxel $x_{s,j}$ in a training image *s*, the weight $w(x_i, x_{s,j})$, can be calculated by a nonlocal means filter as:

$$w(x_i, x_{s,j}) = e^{-\frac{\| p(x_i) - p(x_{s,j}) \|_2^2}{h^2}} \tag{5}$$

where $p(x_i)$ is a cubic patch centered at voxel x_i in the test image, $p(x_{s,j})$ is a cubic patch centered at voxel $x_{s,j}$ in the $s-$th training image, and $\|.\|_2$ is a normalized intensity distance between two patches, and h^2 is a controlling parameter which can be set to the minimum of $\| p(x_i) - p(x_{s,j}) \|^2$ of all selected training patches $p(x_{s,j})$ for a given test patch $p(x_i)$ (Coupé et al., 2011).

To simplify computation, not all patches $p(x_{s,j})$ need to be compared to $p(x_i)$. We define $V_{s,i}$ as a cubic neighbourhood in the training image, centered at the location corresponding to position x_i. Only the patches centered on the voxels $x_{s,j}$ that are part of the *search window* $V_{s,i}$ are considered.

The training patch centered at each voxel $x_{s,j}$ is further considered in a pre-selection process for weight calculation. Basically, to further improve computational efficiency, all patches are pre-selected before calculating the weights to discard the patches whose mean and variation are far away (in terms of intensity) from the test-patch. The pre-selection uses the structural similarity measure (SSIM) (Wang et al., 2004) and can be defined as:

$$SSIM = \frac{2\mu_i \mu_{s,j}}{\mu_i^2 + \mu_{s,j}^2} \cdot \frac{2\sigma_i \sigma_{s,j}}{\sigma_i^2 + \sigma_{s,j}^2} \tag{6}$$

where excluding the subscripts, μ is the mean and σ is the standard deviation of patches $p(x_i)$ and $p(x_{s,j})$.

The final label of the voxel x_i, denoted by $L_T(x_i)$, is a weighted average of all labeled samples inside the search windows around voxels $\{ x_{s,j},\ j \in V_{s,i}, s = 1,2,\ldots, M \}$ from M training images, i.e.,

$$L_T(x_i) = \frac{\sum_{s=1}^{M} \sum_{j \in V_{s,i}} w(x_i, x_{s,j}) L_s(x_{s,j})}{\sum_{s=1}^{M} \sum_{j \in V_{s,i}} w(x_i, x_{s,j})} \tag{7}$$

where $x_{s,j}$ is the *j*-th voxel in the search window $V_{s,i}$, $L_s(x_{s,j})$ is the manual label for voxel $x_{s,j}$, and $w(x_i, x_{s,j})$ is the weight assigned to a pair of patches: the test patch $p(x_i)$ and the training patch $p(x_{s,j})$.

2.3 Combining Appearance Modeling and Nonlocal Means Patch in Segmentation

Although AAM-based segmentation may be good at capturing the global shape variation, it might not be sensitive enough to account for small local shape changes. The local details of the image might be blurred because of the limited training data size, the limited number of eigenvectors derived from the training data, and the limitation of using the linear span of eigenvectors to capture variations. Also, the AAM is not able to generate the local geometrical variation that does not exist in the training data. Motivated by the concept of patch-based label fusion taking advantage of anatomical pattern similarity (Coupé et al., 2011), we combine these two methods into a two-stage segmentation to improve the segmentation accuracy. In this segmentation, the AAM-based segmentation is employed as a first-stage segmentation to identify a coarse contour and its neighbouring area, and the nonlocal means patch-based segmentation is employed as a second-stage segmentation to locally refine the segmentation for voxels in the identified neighbouring area of the coarse contour. The following is a summary of the proposed two-stage segmentation:

- First-stage: Perform AAM-based segmentation and obtain the segmented distance function ϕ. Then, define a local refinement area *R*, namely, the set of voxels inside the distance range [d1, d2] of ϕ.

- Second-stage: For each voxel x_i inside *R*, recalculate the patch similarity function of $\phi_T(x_i)$ using the nonlocal means patch-based refinement method described in Sec. 2.2. Instead of using the manual labels in Eq.7, the signed Euclidean distance functions of the manual labels are integrated into the equation:

$$\phi_T(x_i) = \frac{\sum_{s=1}^{M} \sum_{j \in V_{s,i}} w(x_i, x_{s,j}) \phi_s(x_{s,j})}{\sum_{s=1}^{M} \sum_{j \in V_{s,i}} w(x_i, x_{s,j})} \tag{8}$$

where $\phi_s(x_{s,j})$ is a signed distance function for voxel $x_{s,j}$ in training image s. The distance averaging in the segmentation was also used by Rohlfing and Maurer (2005), where they showed that the distance averaging outperformed the label voting.

- The final segmentation is achieved by thresholding $\phi_T(x_i)$.

3 Illustrative Experiments and Results

The proposed two-stage segmentation algorithm was applied to segment human MTL structures in high-resolution MR images. Two datasets were used for the experiments, with one being a subset of the other. The first dataset comprised 152 healthy adults from 18 to 35 years of age acquired in the context of the International Consortium for Brain Mapping (ICBM) project (Mazziotta et al., 2001). In the experiments, we applied our method to this dataset to study the volumes of MTL structures in the group of healthy young adults. The second dataset was a subset of the first, and was used to validate the method. Termed here the MTL database, the second dataset comprised the first 54 subjects from the ICBM dataset, as their manual labels were available for use as a reference for validation. All MR data were acquired at the Montreal Neurological Institute on a Philips Gyroscan (Best, Netherlands) 1.5T scanner. The T1-weighted scans were acquired with a three-dimensional (3D) spoiled gradient echo sequence with TR = 18 ms, TE = 10 ms, flip angle = 30°, and resolution of 1 mm^3 voxels.

The manual labels of the MTL structures (HC, AG, ERC, PRC, PHC) of the 54 subjects in the MTL database were identified following the protocol defined by Pruessner et al. (2000, 2002) using the software tool "Display" developed at the Montreal Neurological Institute. The inter- and intra-rater variation of the manual labels were evaluated by intra-class correlations (Shrout and Fleiss, 1979). The inter-rater correlation (left - right hemisphere) was 0.86-0.94 for HC, 0.83-0.84 for AG, 0.93-0.95 for ERC, 0.9-0.92 for PRC, 0.88-0.9 for PHC, while the intra-rater correlation (left - right hemisphere) was 0.91-0.94 for HC, 0.91-0.95 for AG, 0.91-0.96 for ERC, 0.92-0.94 for PRC, 0.91-0.93 for PHC. The automatic segmentation results were compared with these manual labels. The similarity between the two

labels is measured by calculating the Dice kappa (κ) (Dice, 1945) [$\kappa = 2 * (V (M \cap A))/(V (M) + V (A))$], where V is the volume, M is the manual label, and A is the automatically segmented label.

In all experiments below, T1-weighted MR images were used for both AAM-based segmentation and patch-based local refinement. As nonlinear registration could provide a better alignment and help improve the AAM based segmentation performance (Hu et al., 2011), we considered nonlinear registration in the two-stage segmentation. For notational simplicity, we named the space for scanned images as the *native* space, the space for linearly registered images as the *source* space, and the space after nonlinear registration as the *model* space. The ANIMAL-based nonlinear registration (Collins et al., 1995) was employed to transform all shape and gray-scale images from the *source* space to the *model* space. The two-stage segmentation was conducted in the *model* space. The final results were converted back to the *source* space via the inverse nonlinear spatial transformation. It was in the *source* space where the automatic segmentation results were compared with these manual labels. The validation of the proposed method and volume characterization were conducted as follows.

- Validation: The validation was performed on the MTL dataset, a subset of the ICBM dataset. The MTL dataset had 54 subjects. In the validation, 54 subjects were partitioned into 4 groups with 14 subjects in each of the first three groups and 12 subjects in the last group. To test one subject in a given group, 40 subjects from the other three groups were selected to build a set of appearance models for the first-stage segmentation. For the sake of simplicity, the common set of appearance models were used to test each of the remaining subjects in the given group. As for the patched-based local refinement in the second-stage segmentation, however, the best 30 out of 53 subjects (excluding the test subject from 54 subjects) were selected.

- Volume characterization: The volume characterization was conducted on the ICBM dataset, which included the MTL dataset used for validation. If a test subject from the ICBM dataset was not in the MTL dataset, all 54 subjects from the MTL dataset were used in building the set of

appearance models and 30 of 54 subjects were selected for the local patch refinement. Otherwise, the segmentation was done as described in the validation bullet above.

In both the validation and volume characterization, the local refinement area was limited to an area formed by voxels with distance range [-2.5, 2.5] of ϕ, where ϕ was the segmentation resulted from the first-stage AAM segmentation. The threshold for SSIM value in Eq. 6 was set to 0.95 in all experiments. The distance range for ϕ and the SSIM value were empirically selected based on simulations. The effects of the distance range of ϕ and SSIM value on the performance of segmenting HC, AG, EPC, and PHC are shown in Figs 2 and 3. Here, the EPC stands for the ento-/peri-rhinal cortex (EPC = ERC + PRC). From Fig. 2, we can see that the best median kappa values measured from 14 subjects (shown as a red bar for each distance range) were obtained by using distance range [-2.5, 2.5] for ϕ, while from Fig. 3 we can see the procedure is quite stable with SSIM values near 0.95 (from the segmentation of three randomly chosen subjects shown).

3.1 Effect of Patch Size on Segmentation Performance

As mentioned earlier, patch-based local refinement analyzes the local intensity similarity between a test patch and each of training patches and then assigns a weight based on the intensity similarity to each patch pair. Accordingly, patch sizes may directly affect the segmentation performance. To study the impact of different patch sizes on segmentation accuracy, we segmented the HC, the EPC, and the PHC using different patch sizes. The κ results of 14 test subjects using different patch sizes are presented in Fig. 4. From the figure, we can see that the best median κ values are with a patch size of 7×7×7 for all structures. The median κ values using $5\times5\times5$ neighbourhood are very close to those from $7\times7\times7$ but the latter are slightly better. These results indicate that a too-small patch size might not be able to capture the local geometry, while a too-big patch size might fail to find the best matched patches in the training data. In the experiments that follow, a patch size of $7\times7\times7$ is used.

3.2 Effect of Search Window Size on Segmentation Performance

As mentioned in Sec.2.2, for a given voxel requiring a local refinement, a cubic neighbourhood in each training image is defined to search for training patches. The cubic neighbourhood size is also known as a search window size. The impact of different search window sizes on segmentation accuracy was also analyzed for the HC, EPC, and PHC. The κ values of 14 test subjects are given in Fig. 5. The results show that the best median κ values are with a search window size of 5×5×5. The κ values from 7×7×7 are shown very close to those from 5×5×5 but the latter are slightly better. The search window size of 5×5×5 (with the best performance here) is slightly smaller than the size of 7×7×7 chosen by Coupé et al. (2011). A possible explanation is that Coupé et al. (2011) used linear registration, while we used a nonlinear registration, which was considered capable of providing a better alignment. In other words, we think a better alignment among subjects can help reduce the search window size. In the following, the search window size is set to 5×5×5.

3.3 Validation of Segmentation Accuracy on MTL Structures

We used the proposed two-stage segmentation method to segment both left and right HC, AG, EPC, and PHC from the MTL dataset of 54 subjects using a leave-one-out method. Table 2 shows the segmentation performance in terms of κ values for the AAM-based method alone, the patch-based method alone, and the proposed combined AAM-based segmentation and patch-based local correction. These experiments show the following:

- For all MTL structures, the mean κ values from the combined AAM and patch-based method are higher than those from the AAM-based method alone or the patch-based method alone, indicating a combination of the global shape constraints from the AAM and the sensitivity to local geometrical change from the patch-based local refinement improves the segmentation accuracy.
- The mixed-factor model repeated measure analysis using multivariate analysis of variance (MANOVA) (Cochran and Cox, 1957) shows a statistically significant effect on κ for all three

segmentation methods ($p < .001$). To further analyze the difference between any two methods, a matched-pair post-hoc t-test was applied. The corresponding p-values are shown in Table 3. Here we refer to being statistically significant as $p < .05$. When the AAM- and patch-based methods were compared, the patch-based method provided better results in segmenting the HC and EPC (see Table 2). The difference in κ between the two methods is statistically significant. For AG segmentation, the AAM-based method provided better results than the patch-based method (see Table 2), and the difference in κ is also statistically significant. There is no statistically significant difference for PHC segmentation ($p = .282$ for left PHC and $p = .805$ for right PHC), although the mean κ for the patch-based method is slightly higher than that for the AAM-based method as shown in Table 2. For all MTL structures, the mean κ values for the combined AAM and patch-based method are higher than those from either the AAM- or patch-based method, and the differences are statistically significant ($p < .05$).

The overlap between the automatic segmented labels and manual labels was also evaluated with a Jaccard index, shown in Table 4. Note that the Dice kappa (κ) and Jaccard (J) index are directly related, i.e., $J = \kappa / (2-\kappa)$ (Shattuck et al., 2001). When the overlap is perfect, both Dice kappa and Jaccard index will be 1.0. When there is an overlap discrepancy, the discrepancy will be mapped to a larger dynamic range in the Jaccard index as compared with the Dice kappa, suggesting that the Jaccard index is more sensitive to the overlap discrepancy.

The improvement in segmentation accuracy of the two-stage combined AAM and patch-based segmentation method can be also observed in Fig. 6, where three example segmentations on the structures of interest are shown and the corresponding κ values provided by the two-stage segmentation are higher than other two automatic segmentation methods. If we further compare the automatic results with corresponding manual labels, the automatic labels are somewhat smoother than the manual ones.

One might be interested in the cases where the two-stage segmentation results did not match well with the manual labels. Two examples on the segmentation of the left HC are shown in Fig. 7, where there are two rows (one for each example) showing the segmentation on 4 sagittal slices. In the upper row (example#1), we can observe an obvious mismatch between the automatically segmented contour and the manual contour in both slice#1 and #2 at the medial border of the HC. In the lower row (example#2), discrepancies can be observed between the automatically segmented contour and the manual contour at the bottom-right corner of the HC. These might be due to the low tissue contrast along the structure boundary, which makes the segmentation difficult.

As a further check on the two-stage segmentation results, we estimated the linear regression on V(A) and V(M), where V is the volume, M is the manual label, and A is the automatically segmented label (from the two-stage segmentation). We also calculate the intra-class correlations (ICC) between V(A) and V(M) as a second measure on how V(A) and V(M) resemble each other. The linear regression results together with R^2 (R square) and ICC values are shown in Figs 8 and 9 for AG, HC, EPC, and PHC, for both left and right sides. We can see R^2 = 0.890 and 0.850 for the left and right AG, 0.907 and 0.927 for the HC, 0.806 and 0.867 for the EPC, and 0.780 and 0.762 for PHC, respectively. The ICC values are 0.904 and 0.885 for the left and right AG, 0.936 and 0.944 for the HC, 0.879 and 0.921 for the EPC, and 0.839 and 0.818 for PHC, respectively. These values indicate extremely good agreement between automatic and manual labels for AG and HC, very good agreement for EPC and good agreement for PHC. Note that the slope of linear regression models is not exactly equal to 1.0. There appears to be a slight over-estimation of smaller structures, and a slight under-estimation of larger structures that may correspond to a regression to the mean. To further check if there is a bias between the automatic volumes (from the two-stage segmentation) and manual volumes, a paired t-test was performed for each MTL structure, and results are listed in Table 5. Generally speaking, for each MTL structure, the mean volumes from the two-stage segmentation is slightly bigger than that of the manual labels, but the volume difference from the paired t-test is not statistical significant ($p > .05$).

Further experiments to check the segmentation speed were performed and the results showed that the proposed two-stage segmentation method (AAM + patch-based method) was able to quickly segment a new subject due to the fact that the first-stage AAM-based segmentation greatly reduced the local refinement area for the second-stage patched-based refinement. To be specific, a rough bounding box around the HC represents a volume of 90000 (=30 x 60 x 50) voxels per image and the number of voxels in the border search region on average is found to be around 9000 voxels per image, a reduction of 90% voxels that represents an equivalent reduction in computational expense for the patch-based segmentation step. The detailed execution time of each step in the training and execution phases of the proposed method are compared with the pure patch-based technique and are shown in Table 6. Since the patch-based method of Coupé et al. (2011) used only linear registration, we included the execution time for linear pre-alignment as well. From Table 6, we can see that with nonlinear pre-alignment, the runtime is ~7.5 minutes for the proposed AAM+Patch method and ~16 minutes for the patch-based method, while with linear pre-alignment, the runtime is ~1.5 minutes for the proposed method and ~10 minutes for the patch-based method. When segmenting a new subject, the overall runtime of the proposed method is more than 50% faster when using nonlinear registration for subject pre-alignment, and 80% faster when using only linear registration, compared to the pure patch-based method (Coupé et al. 2011).

Since the run time reduction in the linear registration case is significant, one might be interested in the segmentation performance with linear registration (Here, by "linear registration", we mean both training and segmentation are done in the linear space). As a check, we tabulate in Table 7 the κ values of segmenting the HC by all three aforementioned methods with only linear registration. The κ values from nonlinear registration are also listed for comparison. The results show that in the linear registration case, both patch-based method and the combined method have similar segmentation performance and each method can provide a significant performance improvement over the AAM method as the κ values are raised to ~0.85 (for patch-based or combined method) from ~0.75 (for the AAM method). On the other hand, the use of nonlinear registration can help increase the κ values, especially for the AAM method, due

to the fact that the nonlinear registration can offer a better structure alignment, which may render the eigen-decomposition analysis used in the AAM method better. As for the patch-based method and the combined method, their performance is also enhanced in the nonlinear registration case. Overall, the segmentation performance with nonlinear registration is found better than that with linear registration. Thus, we will continue using nonlinear registration in our segmentation.

3.4 Volume Analysis of Medial Temporal Lobe Structures in Healthy Young Adults

The integrity of MTL structures is considered as an important marker in the onset and progression of many neurological and neurodegenerative diseases, including Alzheimer's disease and temporal lobe epilepsy. Analyzing the volumetric characteristics of MTL structures in a normal population can thus contribute to a better understanding of the neuropathological changes that may characterize these diseases, and in distinguishing patients from healthy individuals in the early stage of a disease. In this experiment, we used the MTL database with existing MTL segmentations (54 subjects) as training data to segment the MTL structures of 152 subjects in the full ICBM database; more specifically using a leave-one-out method for the first 54 subjects and then the full library of training data for the remaining 98 subjects.

The mean volumes of MTL structures of 152 healthy young adults from the automatic segmentation are summarized in Table 8. Statistical analysis revealed a significantly larger right HC volume ($p = .013$) and a significantly larger left PHC volume ($p < .001$), but no significant difference in AG volume. Apart from the above findings, we found that the left EPC was significantly larger than the corresponding right side ($p = .005$).

No statistically significant differences were found for the HC, AG, and EPC in terms of gender for both left and right hemispheric volumes ($p > .2$) after stereotaxic normalization. The left PHC was significantly larger in females (2,480 mm^3) than in males (2,284 mm^3), but there was no significant difference for the right PHC.

Statistical analysis on the volume of MTL structures against age and gender was further performed using MANOVA and the resulting r and p-values are shown in Table 9. The results indicate the following:

- In females, the left AG volume is weakly positively correlated with age ($r = .249$, $F = 5.309$, $p = .025$), and the right EPC volume is strongly positively correlated with age ($r = .648$, $F = 10.610$, $p = .002$). No significant interaction effect between age and volume in females was observed for other MTL structures.
- In males, the left EPC volume is weakly positively correlated with age ($r = .228$, $F = 5.176$, $p = .026$). No significant interaction effect between age and volume in males was observed for other MTL structures.

4 Discussions and Conclusions

In this paper, we present a novel segmentation algorithm that combines appearance modeling and nonlocal means patch-based local refinement into a general two-stage segmentation framework. During segmentation, the first-stage appearance modeling is used to capture the global shape variation, and the second-stage nonlocal means patch-based method is used to improve the local fitting of the segmentation result. The proposed method was applied to segment the cortical structures of the medial temporal lobes in healthy young adults, and the experimental results demonstrated the feasibility, good performance, and robustness of this algorithm in 3D image segmentation.

As demonstrated in the experimental results, the proposed combination of the AAM and patch-based local refinement did improve the segmentation accuracy in comparison to either the AAM-based method or the patch-based method alone. In addition, the proposed method is able to quickly complete the segmentation of individual subjects. Once the data are aligned (6 minutes per subject for nonlinear registration), only 1 minute is required to process the training data from 54 subjects in 3D with an image size of $70 \times 120 \times 70$ voxels to cover the volume of interest. Segmentation of a new subject requires a

total of ~7.5 minutes (6 minutes for nonlinear registration, less than 30 s for AAM, and ~1 minutes for patch-based local correction using the best 30 templates selected from the 54 training subjects) on a 1.5 GHz Linux PC. This result is much faster than the label fusion procedure proposed by Collins and Pruessner (2010) and the patch-based segmentation by Coupé et al. (2011).

We validated the proposed method using a subset of the ICBM database comprising 54 healthy young adults. The leave-one-out experiments demonstrated the segmentation accuracy of the combined AAM and patch-based methods (mean κ of 0.87 for HC, 0.81 for AG, 0.73 for EPC, and 0.73 for PHC). In order to apply this technique to a different study population such as very young pediatric subjects, very old healthy aging or to disease populations such as epilepsy or Alzheimer's disease where the medial temporal lobe structures are affected, it may be necessary to extend the training library to include subjects from the population to be studied. This way, the shape space spanned by the principal components will better cover the range of the population studied. Furthermore, these new template examples will better represent the intensities used in the patch-based refinement step. Despite these limitations, the procedure presented here enables automatic segmentation of medial temporal lobe structures in the normal population, and thus is applicable to many structure-functional studies where such segmentations are needed.

We like to further point out that in our two-stage segmentation method, the first-stage segmentation can impose a global model constraint on the local refinement area for the second-stage to perform a fine local label fusion. This constrained local refinement area greatly reduces the number of voxels requiring the local refinement as otherwise the local refinement for a large set of voxels would be needed and the resulting computational complexity would be extremely high. On the other hand, this constrained local refinement area may also help the segmentation to perform robustly in regions with low tissue contrast in adjacent structures. To be specific, the local label fusion explores intensity change patterns of patches as a patch inside a segment contour of interest may exhibit an intensity change pattern different from a patch outside of the segment contour. The different intensity change patterns can help assign weights to training

patches for a given test patch according to Eq.5 either in favour of the training patches inside their corresponding contours or in favour of those outside of their corresponding contours. With that, the weighted average for the final segmentation can better determine whether the central voxel of the test patch should be placed inside or outside of the contour. In regions with low tissue contrast in adjacent structures, the intensity change patterns may be homogeneous regardless of the training patches being inside or outside of their segment contours, and the performance of the local label fusion may degrade. In that case, a constrained local refinement area can limit the area where the local label fusion may perform poorly to avoid potential segmentation performance degradation.

The above argument is partly supported by the results in Table 2, where as compared with the patch-based method, the combined method can provide comparable performance for the HC but a larger improvement for the AG, EPC, and PHC. In other words, for structure boundaries with low tissue contrast in certain regions, such as for the AG, EPC, and PHC, the imposed global constraints in terms of a limited local refinement area along the coarse contour identified by the first-stage segmentation may help limit the low-tissue contrast structure boundary area where the local label fusion may not perform well. For the HC, whose structure boundary area in general have a high tissue contrast, the constrained local refinement area might not help too much on the segmentation performance but it definitely helps reduce significantly the computational complexity as the area requiring a local refinement is greatly reduced.

Direct comparison between our technique and others in the literature is difficult because of differences in the anatomical definitions of the structures of interest, types of input data, and quality of manual segmentations. Still, our results are among the best of previous publications for the HC and AG (for details, see Table 1). In particular,

- For HC segmentation, some recently published methods (Klemencic et al., 2004; Chupin et al., 2007; Lijn et al., 2008; Morra et al., 2008; Morey et al., 2009; Aljabar et al., 2009) reported a κ value greater than 0.8. Even more recently, several methods (Collins and Pruessner, 2010; Coupé et al., 2011; Wang et al., 2011) used template warping and label fusion to achieve a high κ value

of greater than 0.88. Our method yielded a mean κ of 0.87 for the HC, which is comparable to the results of those published methods.

- For AG segmentation, most recent methods (Heckemann et al., 2006; Chupin et al., 2007; Morey et al., 2009; Aljabar et al., 2009; Lotjonen et al., 2010; Babalola et al., 2009; Patenaude et al., 2011) reported a κ value of below or equal to 0.8. Only two methods (Collins and Pruessner, 2010; Sabuncu et al., 2010) based on the label fusion technique achieved a κ value of around 0.82. Our method obtained a mean κ of around 0.81 for the AG, with a significant improvement in speed of segmentation over that from Collins and Pruessner (2010).
- As for other MTL structures, there are no published results available with which to compare our findings.

Besides the κ values, in this paper, we also provided the Jaccard index for the segmentation of the AG, HC, EPC and PHC. For some MTL structures, the Jaccard index values were reported at 0.796 for HC and 0.703 for AG by Collins and Pruessner (2010), where an atlas-based label fusion technique was used. Our Jaccard index values for the HC and AG shown in Table 4 are on average ~0.02 worse than those reported by Collins and Pruessner (2010) but our method has a much shorter segmentation runtime. In particular, the atlas-based label fusion technique used 11 atlases for the label fusion procedure and if we consider 6 minutes for nonlinear registration per atlas, the resulting runtime would be 6 x 11 = 66 minutes, while our runtime is 7.5 minutes as discussed earlier.

The structure volumes reported here are slightly different from those published previously by Pruessner (2000, 2001, 2002). This is due in part to the varying numbers of subjects. In Pruessner (2000), a manual segmentation protocol for HC and AG was defined and applied to 40 subjects (20 male and 20 female) from the ICBM dataset acquired at the MNI. In Pruessner (2001), this protocol was used to identify the HC and AG in 80 subjects from the ICBM dataset, selected to match for age and gender. In Pruessner (2002), a new manual segmentation protocol was defined for the temporopolar cortex, PRC,

ERC and PHC, and used to identify these structures on the same set of 40 subjects used in the Pruessner (2000) paper. Here we applied our automated technique to 54 subjects from the ICBM dataset acquired at the MNI, where 40 of these subjects are the same as those used in the Pruessner (2000) and (2002) papers.

Overall, the automatic volumes here are slightly smaller than those previously published manual volumes, but this is not significant for HC or AG. As was the case for the manual labels (Pruessner 2000, 2001), the automatic labels found here showed that the left HC was smaller than the right, and that there was no difference between left and right AG. The same observation on a smaller left HC was also reported by other researchers (Watson et al. 1992; Hasboun et al. 1996; Kidron et al. 1997; Mori et al. 1997) for healthy adults, and patients with Alzheimer's disease. Note that however, for patients with epilepsy, Ashtari et al. (1991), and Cook et al. (1992) reported the left HC being bigger than the right. As for the AG, the finding of no hemispheric differences has been reported by many researchers (Soininen et al. 1994; Mori et al. 1997; Strakowski et al., 1999). However, Watson et al. (1992) found that the right AG was slightly bigger than the left side. While no dependency on age was found for AG for either males or females by Pruessner et al. (2001), we find a slight increase of left AG volume with age for women ($r = .249$, $p = .025$) probably due to the increased number of subjects. Pruessner (2001) found that age was negatively correlated with HC volume in men. We did not find any significant associations with age for HC for either men or women, allowing to speculate whether the eighty randomly chosen subjects studied by Pruessner et al. (2001) had systematic characteristics that created the previously reported age correlation.

For the automatic segmentation of the PHC, as was the case for the manual labels (Pruessner 2002), the automatic results showed that the left PHC was bigger than the right. Although no sex difference was reported for ERC, PHC and PRC in the manual labels of 40 subjects (Pruessner 2002), the automatic results of 152 subjects showed that women had the larger left PHC than men ($p = .019$). On the other hand, for the PRC, Pruessner (2002) found that age was positively correlated with the right PRC volume in women when the volume was not corrected by the collateral sulcus (CS). In our case, we segmented

ERC and PRC jointly and termed it EPC. We found that the left EPC was bigger than the right. Also, the results showed that in women the volume of the right EPC was strongly positively correlated with age ($r = .648$, $F = 10.610$, $p = .002$); while in men the volume of the left EPC was weakly correlated with age ($r = .228$, $F = 5.176$, $p = .026$). These positive age correlations observed in manual and automatic segmentations only in women are intriguing as they perhaps point to a systematic sex difference that warrants further investigation. Their inconsistent appearance with regard to substructure and hemisphere however prevents any firm conclusion at this point in time.

Taken together, the above results provide the impetus for future studies in which the two-stage segmentation method could be routinely applied to MR data from various populations to investigate the association of these structures with various clinical and neuropsychological parameters.

5 References

Aljabar, P., Heckemann, R., Hammers, A., Hajnal, J. V., and Rueckert, D. (2009). Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. *NeuroImage*, 46:726–738.

Ashtari, M., Barr, W. B., Schaul, N., and Bogerts, B. (1991). Three dimensional fast low angle shot imaging and computerized volume measurement of the hippocampus in patients with chronic epilepsy of the temporal lobe. Am. J. Neuroradiol. 12:941-947.

Babalola, K. O., Patenaude, B., Aljabar, P., Schnabel, J., Kennedy, D., Crum, W., Smith, S., Cootes, T., Jenkinson, M., and Rueckert, D. (2009). An evaluation of four automatic methods of segmenting the subcortical structures in the brain. *NeuroImage*, 47:1435–1447.

Barense, M. D., Bussey, T. J., Lee, A. C., Rogers, T. T., Davies, R. R., Saksida, L. M., Murray, E. A., and Graham, K, S. (2005). Functional specialization in the human medial temporal lobe. *Journal of Neuroscience*, 25:10239–10246.

Baxter, M. G. (2009). Involvement of medial temporal lobe structures in memory and perception. *Neuron*, 61:667–677.

Benavides, G. S., Gómez-Ansónc, B., Sainzd, A., Vivesd, Y., Delfinod, M., and Pea-Casanovaae, J. (2010). Manual validation of freesurfer's automated hippocampal segmentation in normal aging, mild cognitive impairment, and alzheimer disease subjects. *Psychiatry research: NeuroImaging*, 181:219–225.

Bishop, C. A., Jenkinson, M., Andersson, J., Declerck, J., and Merhof, D. (2011). Novel fast marching for automated segmentation of the hippocampus (FMASH): method and validation on clinical data. *NeuroImage*, 55:1009–1019.

Cendes, F., Andermann, F., Gloor, P., Evans, A., Jones-Gotman, M., Watson, C., Melanson, D., Olivier, A., Peters, T., and Lopes-Cendes, I. (1993). MRI volumetric measurement of amygdala and hippocampus in temporal lobe epilepsy. Neurology, 43(4):719–725.

Chupin, M., Mukuna-Bantumbakulu, A. R., Hasboun, D., Bardinet, E., Baillet, S., Kinkingnehun, S., Lemieux, L., Dubois, B., and Gamero, L. (2007). Anatomically constrained region deformation for the automated segmentation of the hippocampus and the amygdala: method and validation on controls and patients with alzheimers disease. *NeuroImage*, 34:996–1019.

Chupin, M., Gérardin, E., Cuingnet, R., Boutet, C., Lemieux, L., Lehéricy, S., Benali, H., Garnero, L., Colliot, O., and the Alzheimer's Disease Neuroimaging Initiative, (2009). Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. Hippocampus, 19:579–587

Cochran, W. C. and Cox, G. M. (1957). *Experimental designs. 2nd Edition*. New York: John Wiley & Sons.

Collins, D. L. and Evans, A. C. (1997). Animal: Validation and applications of nonlinea registrationbased segmentation. *International Journal of Pattern Recognition and Artificial Intelligence*, 11(8):1271–1294.

Collins, D. L., Holmes, C. J., Peters, T. M., and Evans, A. C. (1995). Automatic 3D model-based neuro-anatomical segmentation. *Human Brain Mapping*, 3:190–208.

Collins, D. L. and Pruessner, J. C. (2010). Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting animal with a template library and label. *NeuroImage*, 52:1355–1366.

Cook, M. J., Fish, D. R., Shorvon, S. D., Straughan, K., and Stevens, J. M. (1992). Hippocampal volumetric and morphometric studies in frontal and temporal lobe epilepsy. Brain, 115:1001-1015.

Cootes, T., Edward, G., and Taylor, C. (1998). Active appearance model. In *Proc. European Conf. Computer Vision*, volume 2, pages 484–498.

Cootes, T., Cooper, D. H., Taylor, C., and Graham, J. (1995). Active shape models—their training and application. Comput. Vis. Image understanding, 61(1):38-59.

Coupé, P., Manjon, J. V., Fonov, V., Pruessner, J., Robles, M., and Collins, D. L. (2011). Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. *NeuroImage*, 54(2):940–945.

Coupé, P., Eskildsen, S. F., Manjón, J. V., Fonov, V., Collins, D. L. and the Alzheimer's disease Neuroimaging Initiative (2012). Simultaneous segmentation and grading of anatomical structures for patient's classification: Application to Alzheimer's disease. *NeuroImage*, 59:3736-3747.

Dice, L. (1945). Measures of the amount of ecological association between species. *Ecology*, 26(3):297–302.

Duchesne, S., Pruessner, J. C., and Collins, D. L. (2002). Appearance-based segmentation of medial temporal lobe structures. *NeuroImage*, 17:515–531.

Duzel, E., Schiltz, K., Solbach, T., Peschel, T., Baldeweg, T., Kaufmann, J., Szentkuti, A., and Heinze, H. J. (2005). Hippocampal atrophy in temporal lobe epilepsy is correlated with limbic systems atrophy. *Journal of Neurology*, 253(3):294–300.

Eskildsen, S. F., Coupé, P., Fonov, V., Manjón, J. V., Leung, K. K., Guizard, N., Wassef, S. N., RiisØtergaard, L., Collins, D. L., and The Alzheimer's Disease Neuroimaging Initiative (2012). BEaST: Brain extraction based on nonlocal segmentation technique. *Neuroimage*, 59:2362-2373.

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

Fonov, V. S., Evans, A. C., K. Botteron, C. R. A., McKinstry, R. C., Collins, D. L., and the Brain Development Cooperative Group (2011). Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, 54(1):313–327.

Fox, N. C., Freeborough, P. A., and Rossor, M. N., (1996). Visualization and quantification of rates of atrophy in Alzheimers disease. The Lancet, 348:94–97.

Ghanei, A., Soltanian-Zadeh, H., and Windham, J. P. (1998). Segmentation of the hippocampus from brain MRI using deformable contours. *Comput. Med. Imaging Graph*, 22:203–216.

Hasboun, D., Chantome, M., Zouaoui, A., Sahel, M., Deladoeuille, M., Sourour, N., Duyme, M., Marsault, C., and Dormount, D. (1996). MR determination of hippocampal volume: comparison of three methods. Am. J. Neuroradiol., 17(6):1091-1098.

Haegelen, C., Coupé, P., Fonov, V., Guizard, N., Jannin, P., Morandi, X., and Collins, D. L. (2012). Automated segmentation of basal ganglia and deep brain structures in MRI of Parkinson's disease. *Int. J. Comput. Assist. Radiol. Surg.*, DOI: 10.1007/s11548-012-0675-8.

Heckemann, R. A., Hajnal, J. V., Aljabar, P., Rueckert, D., and Hammers, A. (2006). Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *NeuroImage*, 33(1):115–126.

Hu, S. and Collins, D. L. (2007). Joint level set shape modeling and appearance modeling for brain structure segmentation. *NeuroImage*, 36:672–683.

Hu, S., Coupé, P., Pruessner, J. C., and Collins, D. L. (2011). Appearance-based modeling for segmentation of hippocampus and amygdala using multi-contrast MR imaging. *NeuroImage*, 58(2):549–559.

Jack Jr., C. R., Petersen, R. C., OBrien, P. C., and Tangalos, E. G., (1992). MRI-based hippocampal volumetry in the diagnosis of Alzheimer's disease. *Neurology*, 42(1):183–188.

Jack Jr., C. R., Petersen, R. C., Xu, Y. C., Waring, S., OBrien, P. C., Tangalos, E. G., Smith, G. E., Ivnik, R. J., and Kokmen, E., (1997). Medial temporal atrophy on mri in normal aging and very mild Alzheimer's disease. Neurology, 49(3):786–794.

Khan, A. R., Cherbuin, N.,Wen,W., and Anstey, K. j., and Sachdev, P. (2011). Optimal weights for local multi-atlas fusion using supervised learning and dynamic information (SuperDyn): validation on hippocampus segmentation. *NeuroImage*, 56:126–139.

Kidron, D., Black, S. E., Stanchev, P., Buck, B., Szalai, J. P., Parker, J., and Jolesz, F. (1993). Quantitative MR volumetry in Alzheimer's disease. Neurology, 49:1504-1512.

Klemencic, J., Pluim, J., Viergever, M., Schnack, H., and Valencic, V. (2004). Non-rigid registration based active appearance models for 3D medical image segmentation. *Journal of Imaging Science and Technology*, 48(2):166–171.

LeDoux, J. E. (1989). Cognitive-emotional interactions in the brain. *Cogn. Emot.*, 3:267–289.

Lotjonen, J. M., Wolz, R., Koikkalainen, J. R., Thurfjell, L., Waldemar, G., Soininen, H., and Rueckert, D. (2010). Fast and robust multi-atlas segmentation of brain magnetic resonance images. *NeuroImage*, 49:2352–2365.

Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., Woods, R., Paus, T., Simpson, G., Pike, B., Holmes, C., Collins, L., Thompson, P., MacDonald, D., Iacoboni, M., Schormann, T., Amunts, K., Palomero-Gallagher, N., Geyer, S., Parsons, L., Narr, K., Kabani, N., Goualher, G. L., Boomsma, D., Cannon, T., Kawashima, R., and Mazoyer, B. (2001). A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (ICBM). *Philos Trans R Soc Lond B Biol Sci.*, 356:1293–1322.

Morey, R. A., Petty, C. M., Xu, Y., Hayes, J. P., Wagner II, H. R., Lewis, D. V., LaBar, K. S., Styner, M., and McCarthy, G. (2009). A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. *NeuroImage*, 45:855–866.

Mori, E., Yoneda, Y., Yamashita, H., Hirono, N., Ikeda, M., and Yamadori, A. (1997). Medial temporal structures relate to memory impairment in Alzheimer's disease: an MRI volumetric study. *J. Neurol. Neurosurg. Psychiatry*, 63:214–221.

Morra, J. H., Tu, T., Apostoloba, L. G., Green, A. E., Madsen, Avedissian, C., Madsen, S. K., Parikshak, N., Hua, X., Toga, A. W., Jack Jr., C. R., Weiner, M. W., and Thompson, P. M. (2008). Validation of

a full automated 3D hippocampal segmentation method using subjects with alzheimers disease mild cognitive impairment, and elderly controls. *NeuroImage*, 43(1):59–68.

Patenaude, B., Smith, S. M., Kennedy, D. N., and Jenkinson, M. (2011). A bayesian model of shape and appearance for subcortical brain segmentation. *NeuroImage*, 56(3):907–922.

Powell, S., Magnotta, V. A., Johnson, H., Jammalamadaka, V. K., Andreasen, N. C., and Pierson, R. (2008). Registration and machine learning based automated segmentation of subcortical and cerebellar brain structures. *NeuroImage,* 39(1):238–247.

Pruessner, J. C., Li, L. M., Serles, W., Pressner, M., Collins, D. L., Kabani, N., Lupien, S., and Evans, A. C. (2000). Volumetry of hippocampus and amygdala with high-resolution MRI and three-dimensional analysis software: Minimizing the discrepancies between laboratories. *Cerebral Cortex*, 10:433–442.

Pruessner, J. C., Collins, D. L., Pruessner, M. and Evans, A. C. (2001). Age and gender predict volume decline in the anterior and posterior hippocampus in early adulthood. *J. of Neuroscience,* 21(1):194-200.

Pruessner, J. C., Kohler, S., Crane, J., Pruessner, M., Lord, C., Byrne, A., Kabani, N., Collins, D. L., and Evans, A. C. (2002). Volumetry of temporopolar, perirhinal, entorhinal and parahippocampal cortex from high-resolution MR images: Considering the variability on the collateral sulcus. *Cerebral Cortex*, 12:1342–1353.

Rohlfing, T. and Maurer C. R. Jr., (2005). Shape-based averaging for combination of multiple segmentations. In Proc. of 8th MICCAI, 8(pt 2):838-845.

Rousseau, F., Habas, P., and Studholme, C. (2011). A supervised patch-based approach for human brain labeling. *IEEE Transaction on Medical Imaging,* 30(10):1852-1862.

Sabuncu, M. R., Yeo, B. T., van Leemput, K., Fischl, B., and Golland, P. (2010). A generative model for image segmentation based on label fusion. *IEEE Transaction on Medical Imaging*, 29(10):1714–1729.

Shattuck, D. W., Sandor-Leahy, S.R., Schaper, K.A., Rottenberg, D.A., and Leahy, R.M. (2001). Magnetic resonance image tissue classification using a partial volume model, NeuroImage, 13:856-876.

Shattuck, D. W., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salamon, G., Narr, K., Poldrack, R., Bilder, R., and Toga, A. (2008). Construction of a 3D probabilistic atlas of human cortical structures. *NeuroImage*, 39(3):1064–1080.

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

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

Soininen, H. S., Partanen, K., Pitkanen, A., Vainio, P., Hanninen, T., Hallikainen, M., Koivisto, K., and Riekkinen, P. J. (1994). Volumetric MRI analysis of the amygdala and the hippocampus in subjects with age-associated memory impairment: correlation to visual and verbal memory. Neurology, 44:1660-1668.

Strakowski, S. M., DelBello, M. P., Sax, K. W., Zimmerman, M. E., Shear, P. K., Hawkins, J. M., and Larson, E. R. (1999). Brain magnetic resonance imaging of structural abnormalities in bipolar disorder. Arch. Gen. Psychiat., 56:254-260.

Toth, R., and Madabhushi, A., (2012). Multi-feature landmark-free active appearance models: application to prostate MRI segmentation. IEEE Trans. Med. Imaging, in press.

van der Lijn, F., Heijer, T., Breteler, M. M. B., and Niessen, W. J. (2008). Hippocampus segmentation in MR images using atlas registration, voxel classification, and graph cuts. *NeuroImage*, 43:708–720.

Wang, H., Das, S., Suh, J. W., Altinay, M., Pluta, J., Craige, C., Avants, B., Yushkevich, P. A., and The Alzheimer's Disease Neuroimaging Initiative (2011). A learning-based wrapper method to correct systematic errors in automatic image segmentation: Consistently improved performance in hippocampus, cortex and brain segmentation. *NeuroImage*, 55:968–985.

Wang, H., Suh, J. W., Das, S., Pluta, J., Altinay, M., Yushkevich, P. (2011b). Regression-Based Label Fusion for Multi-Atlas Segmentation. *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, June 2011, 1113-1120.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transaction on Image Processing*, 13:600–612.

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

Xu, Y., Jr., C. J., O´ Brien, P. C., Kokmen, E., Smith, G. E., Ivnik, R. J., Boeve, B. F., Tangalos, R. G., and Peterson, R. C. (2000). Usefulness of mri measures of entorhinal cortex versus hippocampus in AD. *Neurology*, 54(9):1760–1767.

Table 1 Review of segmentation methods

Author	Method summary	Test data	Result in terms of κ		
			HC	AG	PHC
Fischl et al., 2002	FreeSurfer	70 healthy young	0.8	0.75-0.78	
Klemencic et al., 2004	Appearance model		0.8		
Heckemann et al., 2006	Multiatlas based	30 normal	0.82	0.8	0.81*
Chupin et al., 2007	Seeding + morphology	16 healthy young	0.84	0.8	
	region growing	8 AD	0.84	0.76	
Powell et al., 2008	Machine learning	15 subjects	0.85		
	based classification				
van der Lijn et al., 2008	Multiatlas +	20 older adults	0.858		
	graph cuts				
Morra et al., 2008	Auto context model	21 AD	0.835-0.859		
	+adaboost				
Morey et al., 2009	FSL/FIRST	20 healthy	0.79	0.73	
Aljabar et al., 2009	Multiatlas + image	275 subjects	0.84	0.78	
	similarity selection	from 4 to 83			
		years old			
Lotjonen et al., 2010	Multiatlas +	1,000 AD,	0.82-0.88	0.77	
	intensity modeling	MCI and CN			
Babalola et al., 2009	CFL**	270 subjects	0.84	0.78	
	EMS	from 4 to 83	0.77	0.71	
	PAM	years old	0.77	0.67	
	BAM		0.79	0.73	
Collins and Pruessner, 2010	Multiatlas +	80 healthy	0.887	0.826	
	label fusion	young adults			
Benavides et al., 2010	FreeSurfer	41 healthy older adults	0.78		
		23 MCI /AD			
Sabuncu et al., 2010	Label fusion	28 healthy	0.82~0.87	0.8~0.82	
		11 MCI/AD			
Coupé et al., 2011	Nonlocal means	80 healthy	0.884		
		young adults			

Table 1 Review of segmentation methods (cont.)

Author	Method summary	Test data	Result in terms of κ		
			HC	AG	PHC
Bishop et al., 2011	FMASH	CMA: 9 normal and 8 AD BPSA: 16 BP and 16 normal	0.82 0.8		
Patenaude et al., 2011	Bayesian appearance	336 subjects both normal and pathological brain	0.81	0.74	
Khan et al., 2011	Multiatlas + spatially local selection	69 middle-aged 37 older adults	0.833 0.853		
Wang et al., 2011	Multiatlas + error correction	57 normal 82 MCI	0.908 0.893		

* Parahippocampal + ambient gyri;

**CFL: classifier fusion and labeling; EMS: expectation-maximization using a brain atlas; PAM: profile active appearance models; BAM: Bayesian appearance models.

Table 2 Mean κ values from AAM-based, patch-based, and AAM+patch-based methods

	AAM		Patch		AAM + patch	
	Left	Right	Left	Right	Left	Right
HC	0.851 (0.028)	0.862 (0.020)	0.862 (0.028)	0.866 (0.022)	0.867 (0.025)	0.873 (0.019)
AG	0.800 (0.048)	0.792 (0.055)	0.790 (0.048)	0.781 (0.058)	0.812 (0.043)	0.803 (0.053)
EPC	0.711 (0.068)	0.697 (0.083)	0.720 (0.066)	0.703 (0.082)	0.735 (0.066)	0.714 (0.082)
PHC	0.696 (0.067)	0.707 (0.060)	0.691 (0.058)	0.709 (0.052)	0.730 (0.048)	0.739 (0.047)

Notes: Values are mean κ values shown with standard deviations in parentheses.

Table 3 Matched-pair t-test results between different segmentation methods

(threshold for significance $p < .05$)

	AAM vs. Patch		AAM vs. AAM + Patch		Patch vs. AAM + Patch	
	Left	Right	Left	Right	Left	Right
HC	< **0.001**	**0.039**	< **0.001**	< **0.001**	**0.016**	< **0.001**
AG	**0.046**	**0.018**	< **0.001**	< **0.001**	< **0.001**	< **0.001**
EPC	**0.040**	**0.026**	< **0.001**	< **0.001**	< **0.001**	< **0.001**
PHC	0.282	0.805	< **0.001**	< **0.001**	< **0.001**	< **0.001**

Notes: Values are p-values of t-test.

Table 4 Dice Kappa and Jaccard index for MTL segmentation of the combined AAM and patch-based method

	Dice Kappa		Jaccard index	
	Left	Right	Left	Right
AG	0.812 (0.043)	0.803 (0.053)	0.686 (0.058)	0.673 (0.071)
HC	0.867 (0.025)	0.873 (0.019)	0.766 (0.038)	0.775 (0.029)
EPC	0.735 (0.066)	0.714 (0.082)	0.587 (0.080)	0.562 (0.093)
PHC	0.730 (0.048)	0.739 (0.047)	0.578 (0.057)	0.589 (0.057)

Notes: Values are mean κ values shown with standard deviations in parentheses.

Table 5 Statistical analysis on the MR volumes of medial temporal lobe structures (volumes normalized in the stereotaxic space)

	Two-stage segmentation (volume)		Manual segmentation (volume)		Paired t-test (two-stage vs manual) (p-value)	
	Left	Right	Left	Right	Left	Right
AG	1496 (218)	1478 (254)	1448 (298)	1435 (274)	0.086	0.089
HC	3983(486)	4107(556)	3901(528)	4072(596)	0.092	0.103
EPC	3163(720)	3080(810)	3125(848)	3020(926)	0.481	0.314
PHC	2279(461)	2047(348)	2199(419)	1989(350)	0.064	0.057

Notes: Volume values are mean volumes in units of 1 mm^3, with standard deviations in parentheses.

Table 6 Segmentation runtime comparison for the proposed AAM+Patch method and the patch-based technique running on a 1.5 GHz Linux PC.

Processing steps	AAM + patch (Nonlinear registration)	Patch alone (Nonlinear registration)	AAM + patch (Linear registration)	Patch alone (Linear registration)
Training time				
Training image nonlinear registration	6 minutes per subject	6 minutes per subject	0	0
AAM training (40 training subjects)	1.5 minutes	0	1.5 minutes	0
Patch training (pre-calculate mean and variance for each training patch)	2 minutes	2 minutes	2 minutes	2 minutes
Run time				
Test image nonlinear registration	6 minutes	6 minutes	0	0
AAM-based segmentation (Least square solution)	Less than 30 seconds	0	Less than 30 seconds	0
Template selection	1 second	1 second	1 second	1 second
Patch-based refinement	~1 minute (on average ~9000 voxels)	~10 minutes (90000 voxels)	~1 minutes (on average ~9000 voxels)	~10 minutes (90000 voxel)
Inverse nonlinear transformation	3 seconds	3 seconds	0	0
Total runtime	~7.5 minutes	~16 minutes	~1.5 minutes	~10 minutes

Table 7 Segmentation performance comparison in terms of Dice κ for the HC with linear and nonlinear registration

Methods	Registration	Left HC	Right HC
AAM alone (T1 images)	Linear	0.746 (0.061)	0.755 (0.043)
AAM alone (T1 images)	Nonlinear	0.851 (0.028)	0.862 (0.020)
Patch-based segmentation alone	Linear	0.843 (0.050)	0.848 (0.040)
Patch-based segmentation alone	Nonlinear	0.862 (0.028)	0.866 (0.022)
AAM (T1 images) + Patch based method	Linear	0.841 (0.035)	0.845 (0.038)
AAM (T1 images) + Patch based method	Nonlinear	0.867 (0.025)	0.873 (0.019)

Notes: Values are mean κ values shown with standard deviations in parentheses.

Table 8 MR volumetry of medial temporal lobe structures in healthy young adults (volumes normalized in the stereotaxic space)

	Left (volume)	Right (volume)	t-test (Left vs. Right) p-value
AG	1455 (199)	1422 (217)	0.186
HC	3945 (430)	4072 (454)	**0.013**
EPC	3260 (736)	3023 (737)	**0.005**
PHC	2369 (507)	2092 (435)	< **0.001**

Notes: Volume values are mean volumes in units of 1 mm^3, with standard deviations in parentheses.

Table 9 r and p values from statistical analysis on volumes in the stereotaxic space of MTL structures against age according to the gender in 152 healthy young adults

	Females (n=66)		Males (n=86)	
	Left	Right	Left	Right
AG	**0.249 (0.025)**	-0.011 (0.626)	0.084 (0.530)	0.110 (0.181)
HC	0.155 (0.215)	0.194 (0.116)	-0.071 (0.441)	-0.102 (0.324)
EPC	0.104 (0.194)	**0.648 (0.002)**	**0.228 (0.026)**	-0.011 (0.990)
PHC	-0.023 (0.851)	-0.014 (0.914)	0.010 (0.939)	-0.022 (0.842)

Notes: Values are *r* values, with p-values in parentheses.

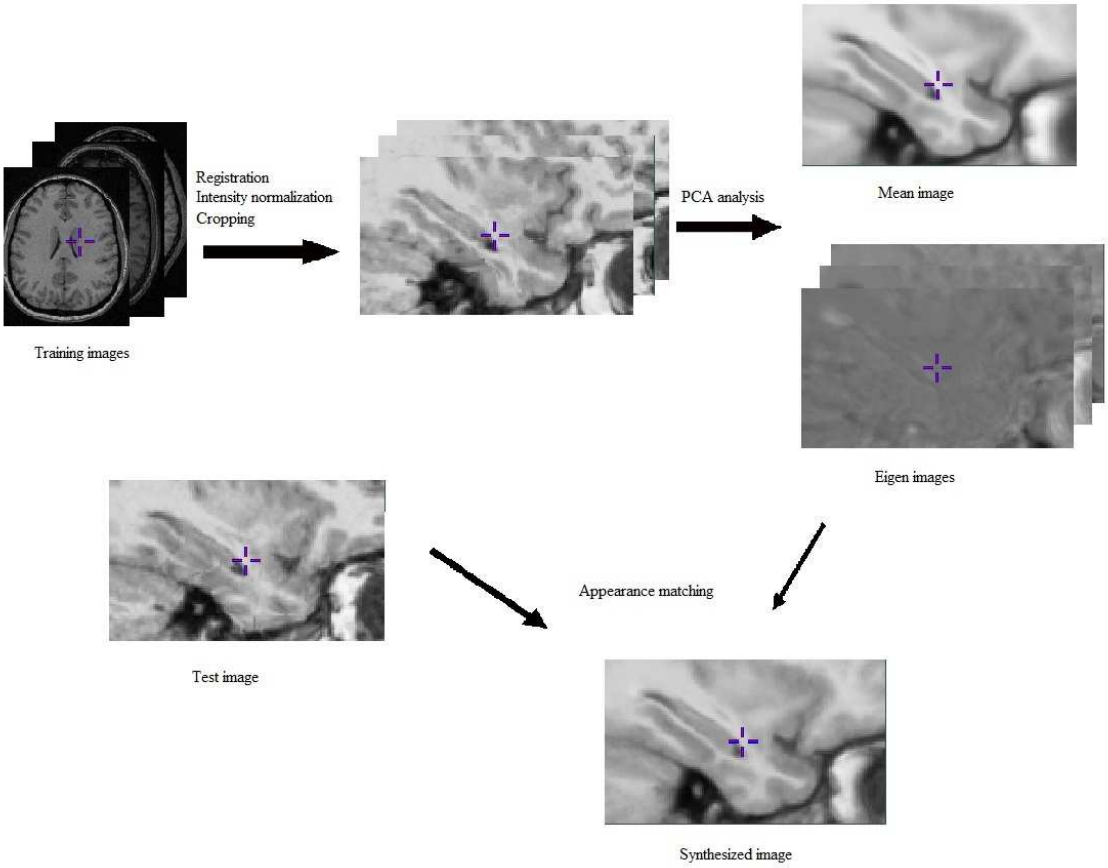

Figure 1: Processing pipeline for the appearance-model based segmentation

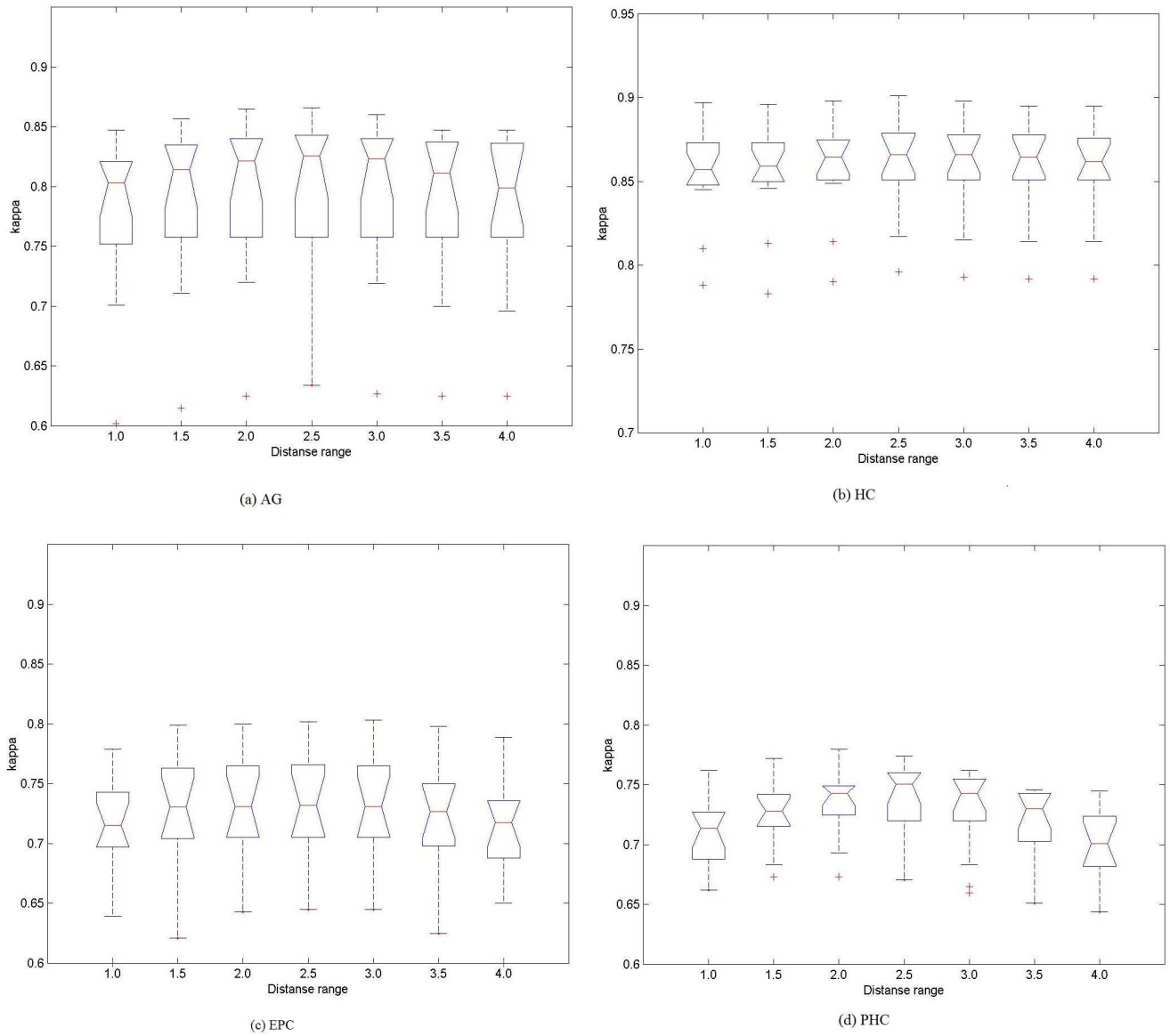

Figure 2: Impact of distance range ±d of ϕ on the performance of the two-stage segmentation. Kappa (κ) values of 14 test subjects for d from 1.0 to 4.0mm with steps of 0.5mm. (a) AG, (b) HC, (c) EPC, and (d) PHC.

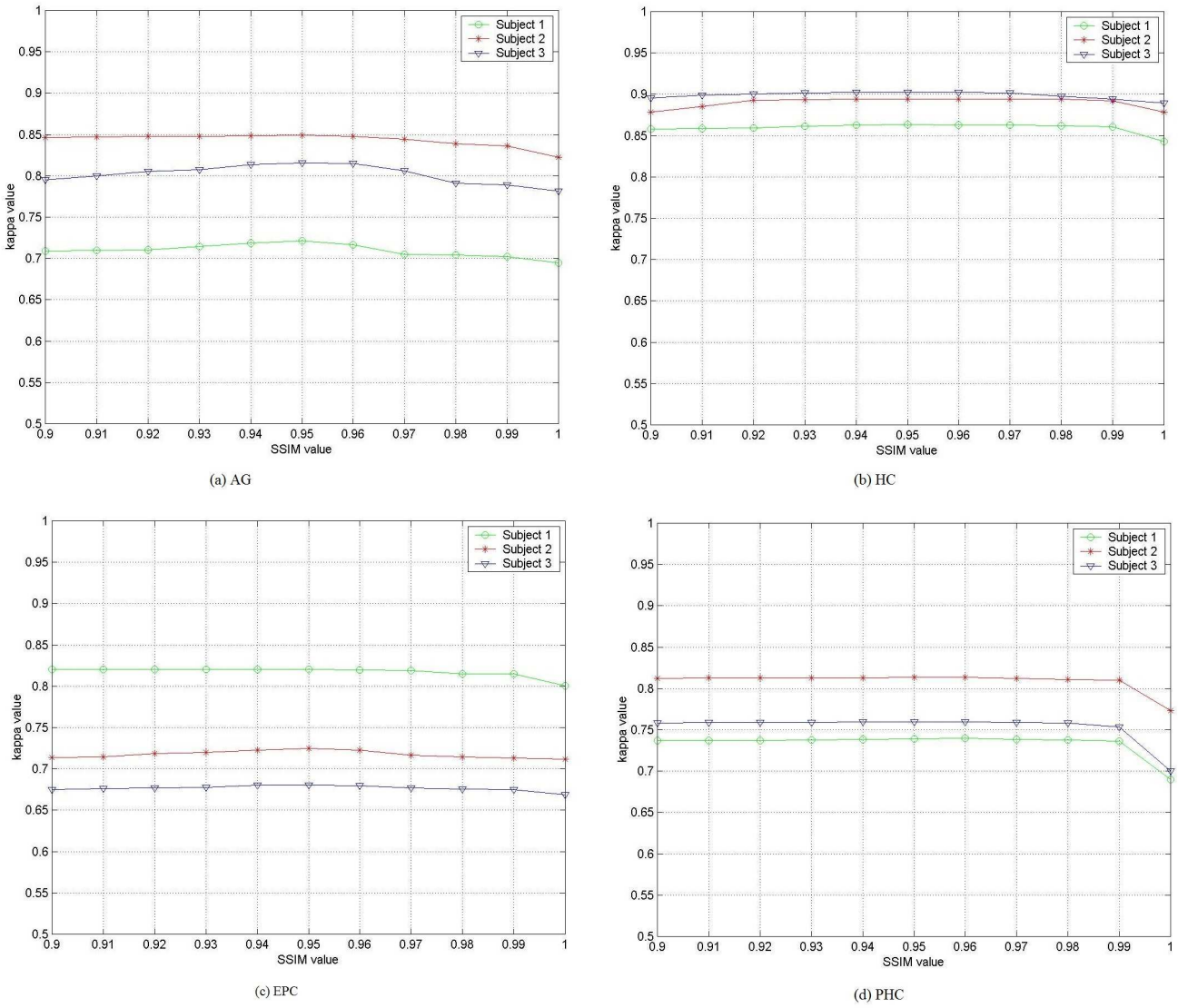

Figure 3: Impact of different SSIM values on the performance of two-stage segmentation. Kappa (κ) values of 3 randomly chosen test subjects for SSIM from 0.9 to 1.0. (a) AG, (b) HC, (c) EPC, and (d) PHC.

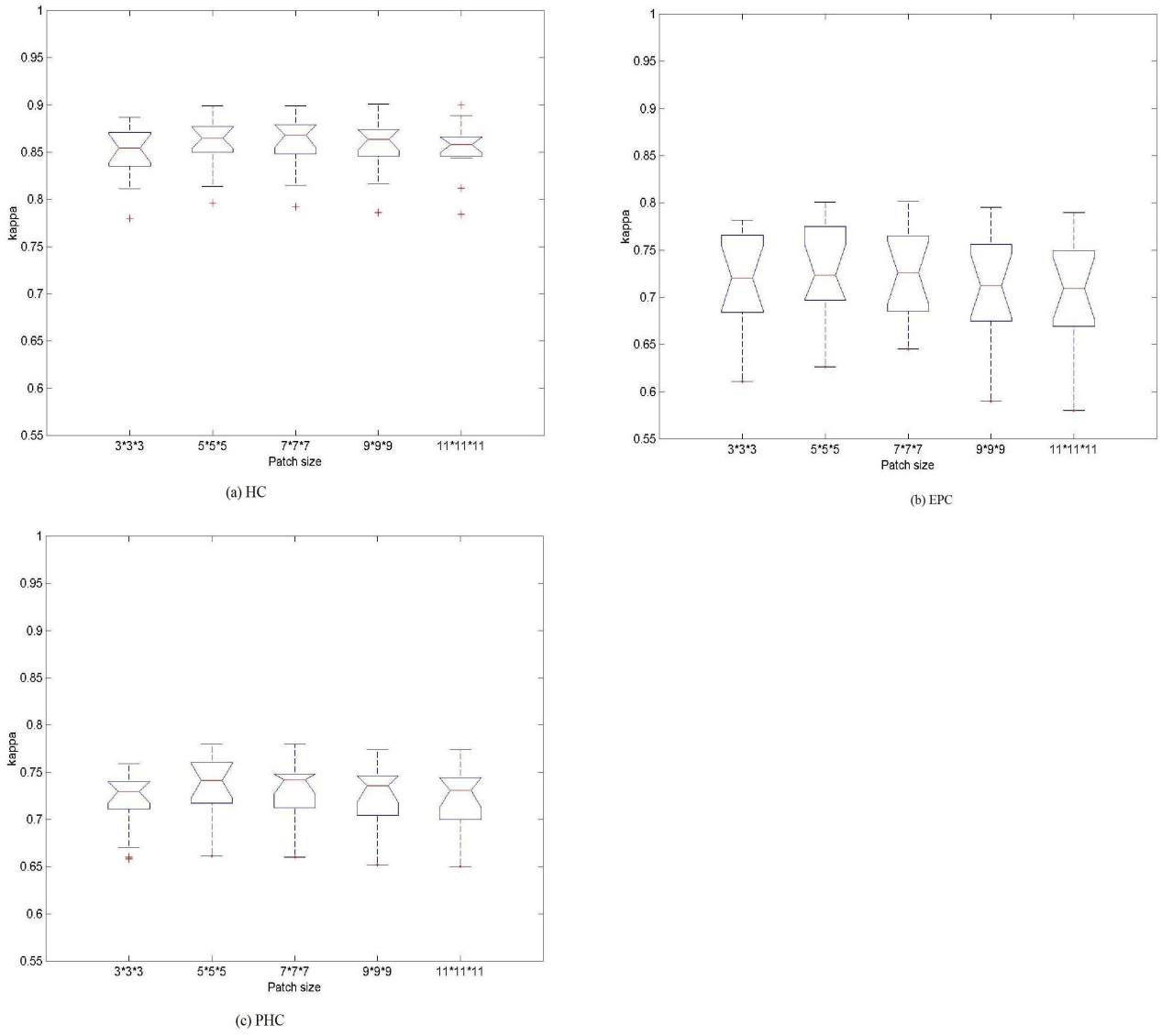

Figure 4: Impact of patch size on segmentation performance. Kappa (κ) values of 14 test subjects under different patch sizes. (a) HC, (b) EPC, and (c) PHC.

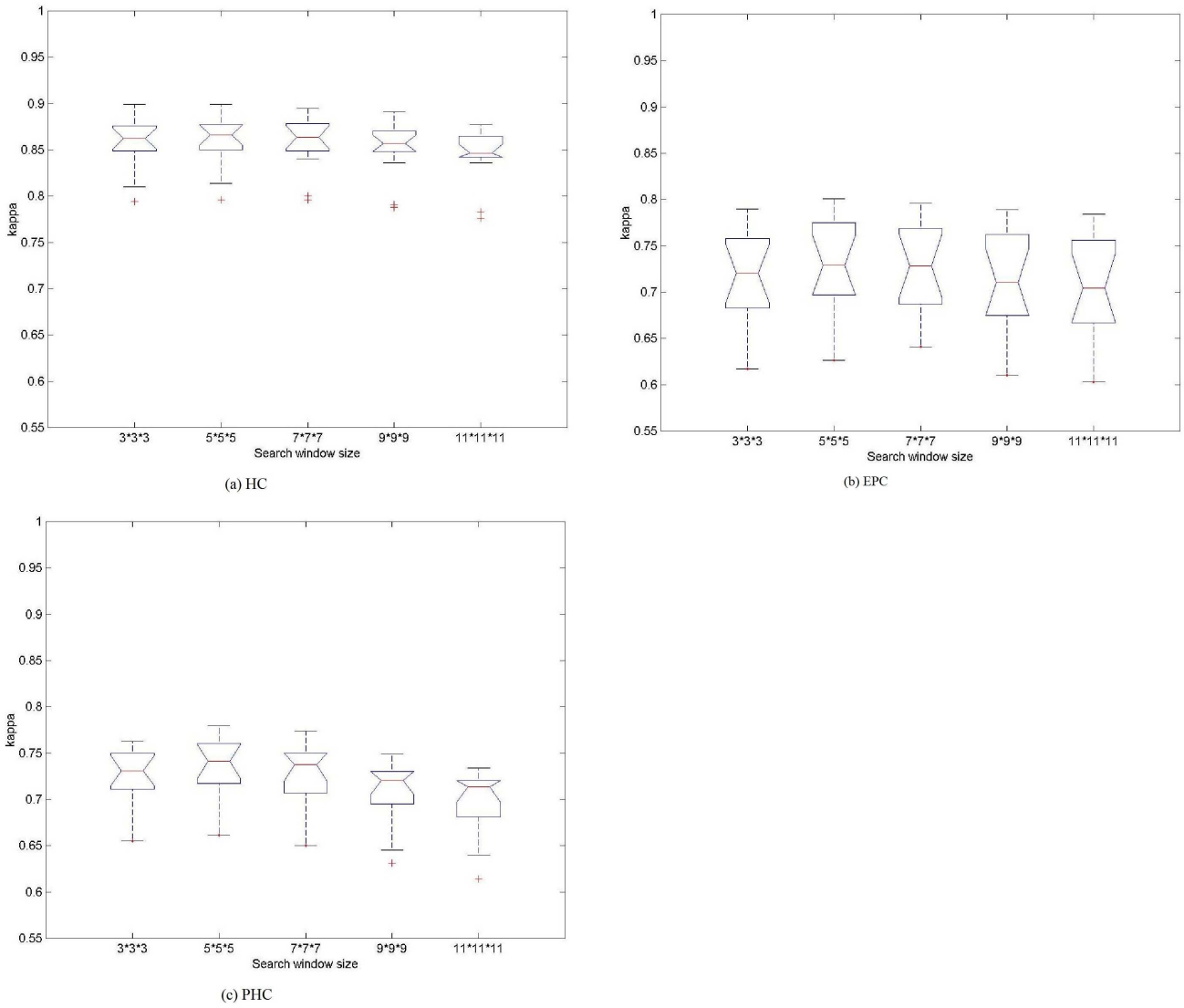

Figure 5: Impact of search window size on segmentation performance. Kappa (κ) values of 14 test subjects under different search window sizes. (a) HC, (b) EPC, and (c) PHC.

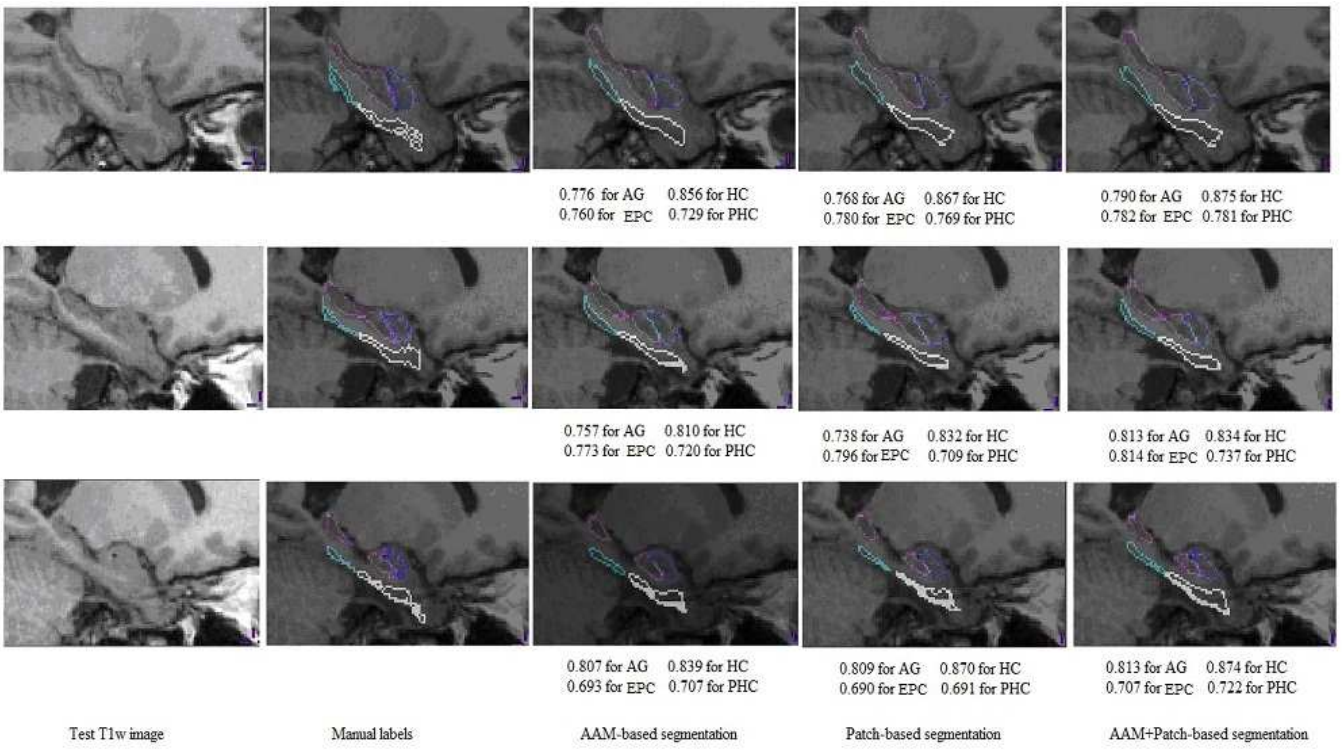

Figure 6: Two-dimensional visualization of 3D segmentation results of three test subjects with average κ value: one test subject per row, and columns from left to right for test image and segmented contours from manual label and three automatic segmentation methods . κ values shown under each graph. The segmented contours of different structures rendered on top of the corresponding T1-weighted test MR image with this color coding: purple for HC, blue for AG, sky blue for EPC, and white for PHC.

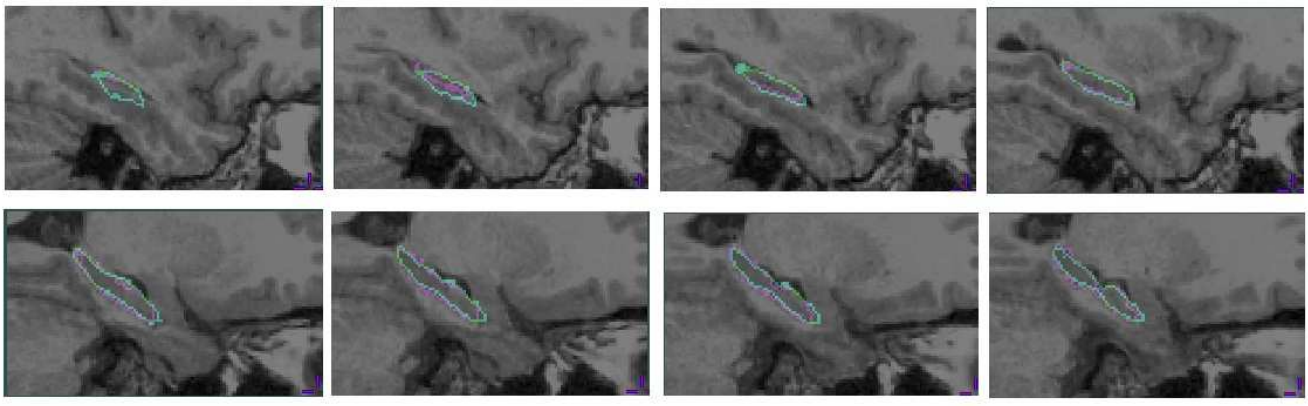

Figure 7: Two examples showing the two-stage segmentation mismatching the true structure boundary of the left HC: One example per row – upper row for example#1, lower row for example#2; Each example shows 4 sagittal slices through the medial temporal lobe. Color coding -- purple for the manual labeled contour; sky blue for the automatically segmented contour.

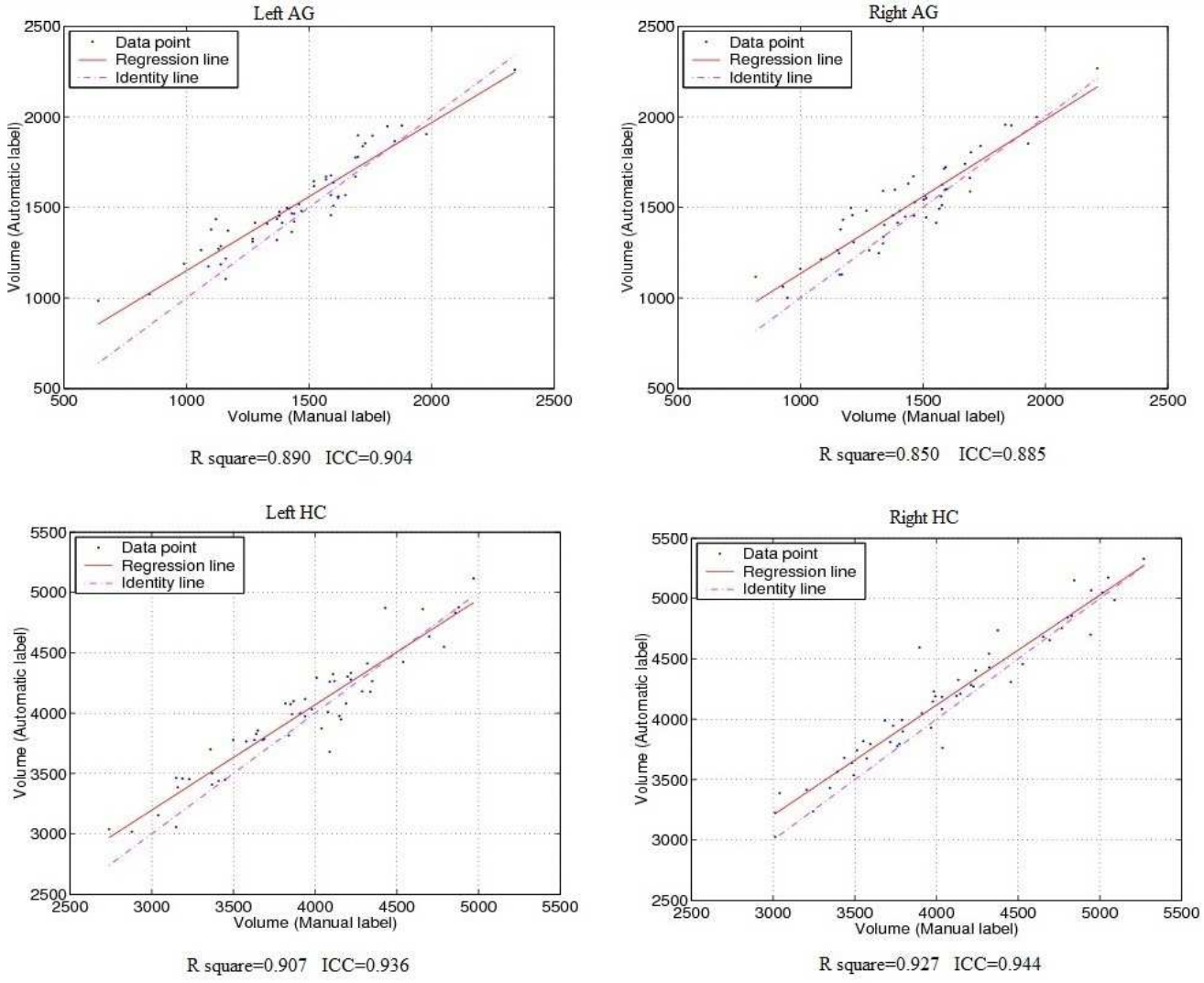

Figure 8: Volumetric comparison between the two-stage segmentation results and manual labels for the HC and AG (volumes normalized in stereotaxic space).

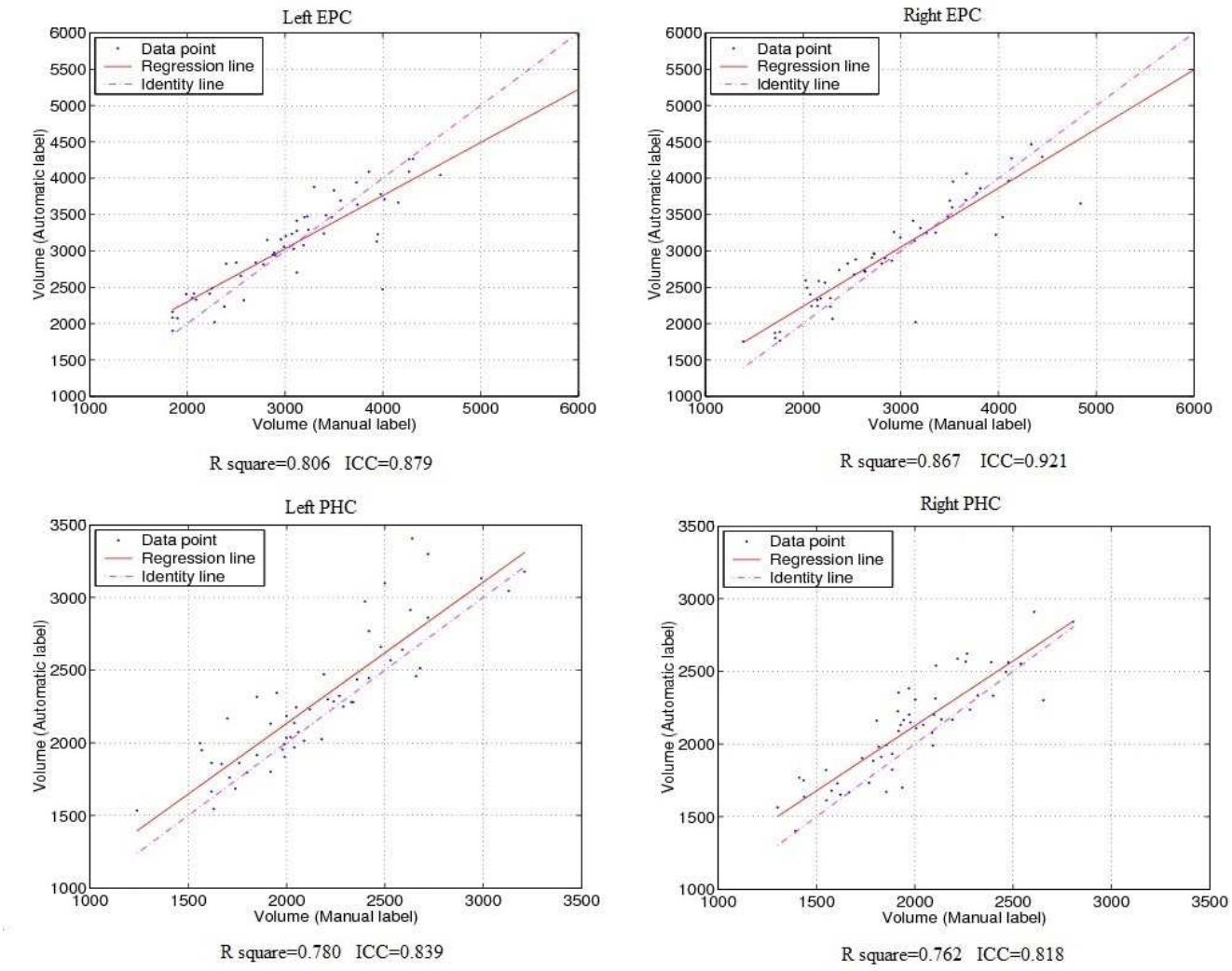

Figure 9: Volumetric comparison between the two-stage segmentation results and manual labels for the EPC and PHC (volumes normalized in stereotaxic space).