A fast approach for hippocampal segmentation from T1-MRI for predicting progression in Alzheimer's disease.

Carlos Platero, M. Carmen Tobar

Health Science Technology Group, Technical University of Madrid, Ronda de Valencia 3, 28012, Madrid, Spain.

Abstract

Background: We provide and evaluate an open-source software solution for automatically measuring hippocampal volume and hippocampal surface roughness based on T1-weighted MRI, which allows for discriminating between patients with Alzheimer's disease (AD) or mild cognitive impairment (MCI) and elderly controls (NC) using only one scan.

New Method: This solution is based on a fast multiple-atlas segmentation technique, which combines a patch-based labeling method with an atlas-warping using non-rigid registrations.

Results: The classifications are comparable to the best classifications in a large clinical dataset. For AD vs control, we obtain a high degree of accuracy, approximately 90%. For MCI vs control, we obtain accuracies ranging from 70% to 78%. The average time for the hippocampal segmentation from a T1-MRI is less than 17 minutes.

Comparison with Existing Method: In this study, we investigate a combination of our method with annotations using the Harmonized Hippocampal Protocol (HarP). We compare its capabilities with the FreeSurfer method and verify its impact on segmentation and diagnostic group separation capabilities. Our approach is developed and validated using 134 subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database with annotations from HarP. Then, this method, tuned with the best parameters, is applied to 162 subjects from a private image database.

Conclusions: Our approach with HarP annotations has a high level of accuracy for segmentation of the hippocampus and is robust to multi-site data. The bio-markers extracted from our proposed method have discriminative

power based on a scalar feature, showing robustness in generalization and avoid overfitting.

Keywords:
Atlas-based segmentation, Alzheimer's Disease, Patch-based label fusion, Hippocampal segmentation, Magnetic resonance imaging

1. Introduction

The analysis of the hippocampus is a very active field of research because it is one of the first structures where early Alzheimer's disease (AD) pathology is observed [1]. Patients suffering from the initial stages of AD are mostly clinically classified as amnestic mild cognitive impairment (MCI), but not all patients with amnestic MCI will develop AD [2]. It has been shown that measurements of the hippocampus from T1-weighted (T1-W) MRI are useful markers of AD and progression of the clinical decline at a MCI stage [2, 3]. A challenge of the neuroimaging is to help in the diagnosis of early AD, particularly in amnestic MCI patients or prodromal AD.

Several approaches have been proposed for automatically classifying patients with AD and/or MCI from hippocampal segmentations [4]. The manual segmentation of the hippocampus is considered the gold standard but is time consuming [5]. Several studies have demonstrated that hippocampal volumetry from the manual segmentation can distinguish patients with AD from controls with a high degree of accuracy (80% to 90%) and that the discrimination between MCI patients and controls is substantially lower (60% to 74%) [6]. The manual segmentation of the hippocampus requires considerable training and consumes more than 1 hour of work. Consequently, many automatic approaches have been proposed for extracting the hippocampal structures from brain MRI [7, 8, 9, 10, 11]. Among such approaches, atlas-based methods have been demonstrated to outperform other algorithms [12] that rely on manual segmentations. However, the hippocampus is a complex anatomical structure, and different manual segmentation protocols have been proposed. In fact, a review by Konrad et al [13] identified 71 hippocampal tracing methods. The absolute volume differences between certain protocols may vary by 30%. The lack of an agreed reference procedure for manual segmentation is a major barrier to the widespread acceptance and use of hippocampal measures for clinical diagnosis. Defined standard operating procedures for hippocampal segmentation are required for its concrete use as

an element to extract bio-markers. Furthermore, the disease status of subjects used in the atlas set may affect the results obtained on a different data set. Most studies have atlases based on normal controls [14, 15, 16, 10, 11]. Atlases should be customized for the pathological studies. An international effort to harmonize existing protocols has defined the Harmonized Hippocampal Protocol (HarP) [17, 18]. This protocol proved to be very reliable and to provide a hippocampal segmentation estimate that can be considered as a standard measure, enabling the use of the hippocampal measures as proper bio-markers for AD and MCI.

After obtaining the segmentation of the hippocampus, several bio-markers have been proposed. Generally, the analysis of the volume and/or shape of this anatomical structure is used. The hippocampal volume using MRI is the main criterion for allowing a diagnosis of AD [2, 16, 9, 4]. This bio-marker enables a separation between AD and normal controls (NC) with an accuracy of approximately $72\% - 74\%$ over the entire Alzheimer's Disease Neuroimaging Initiative (ADNI) database [4]. This limited capability to classify AD patients using only the volume may be due to both a simplification of the atrophy patterns to a global measure and discrepancies caused by the manual protocols. Not only is there interest regarding the hippocampal volume in a unique sample but it has also been observed in measuring volume changes over time [9]. In Wolz et al. [19], the authors reported a correct classification rate of 82% for NC versus AD on 568 images of the ADNI dataset. However, this type of approach requires several scans for a given patient. Additionally, volumetric analyses do not provide information about the precise locations of morphological changes that characterize the appearance and progression of AD.

Recently, several methods for the regional hippocampal shape analysis have been proposed [20, 6, 21] for capturing detailed hippocampal structural modifications to obtain a more accurate classification. In the comparison of approaches with the same images [6], methods based on shape analysis [6] yield slightly better classification than volumetric approaches. Furthermore, shape analysis approaches allow for the identification of regions in the hippocampus between NC and disease groups, which contributes to the prediction of the conversion from MCI to AD [20, 21, 22].

An emerging method is to segment subfields of the hippocampus [23, 24, 25, 26]. This approach appears promising because it is potentially able to detect more detailed atrophic patterns. However, ultra-high resolution MRI is required, which is not yet the standard in clinical practice and thus

currently limits the practical applicability of this approach.

Therefore, the development of new methods capable of estimating subtle anatomical modifications of the hippocampus appears to be critical for obtaining a better classification rate.

In this study, we use an atlas-based segmentation with annotations from the HarP. We compare our approach with the FreeSurfer method [14] and verify its impact on segmentation and diagnostic group separation capabilities. Two bio-markers, namely, the normalized hippocampal volume and the hippocampal surface roughness, are used for evaluating the segmentation algorithms using their capabilities to detect structural changes caused by AD. For this purpose, we used the ADNI database and a particular image database.

The contribution of this paper is a flow-chart for the automated hippocampal segmentation based on multi-atlas segmentation with the HarP. The approach relies on local contexts by patches representation in both intensity and labeling, allowing constraints on the shape prior, which is essential in the acquisition of the hippocampus using T1-MRI due to the low contrast in intensity. Our segmentation results, guided by the HarP, make the bio-markers used for detecting AD more robust compared to other approaches.

2. Materials

In this study, two database were used to validate the proposed approach and compare it to FreeSurfer: (i) a subset of the ADNI database (http://www.loni.ucla.edu/ADNI), which are 134 images that was labeled following the HarP and (ii) a private database from the Laboratory of Cognitive and Computational Neuroscience (LCCN) of the Center of Biomedical Technology (Technical University of Madrid) [27, 28]. We denote this set as LCCN.

For ADNI images, MRI acquisition was performed according to the ADNI acquisition protocol [29]. The set of the ADNI-HarP was also used to construct the atlas set. The identification numbers of the ADNI-HarP images used in this study are reported in (http://www.hippocampal-protocol.net). The subjects included in this study are 44 controls, 45 patients diagnosed with MCI and 45 patients with AD.

LCCN images are acquired using a Fast Spoiled Gradient Echo sequence with the following parameters: TR/TE/TI = 11.2/4.2/450 ms; flip angle of

Table 1: Demographic details of the subset of the ADNI-HarP database used as the atlases with the HarP. Mean (SD) unless specified otherwise.

	NC	I-MCI	II-MCI	AD
Number of subjects	44	16	29	45
Age, years	76.1 (7.4)	73.1 (7.8)	75.2 (8.1)	74.5 (8.1)
Gender male (%)	21 (52%)	8 (50%)	18 (62%)	21 (47%)
MMSE	29.0 (1.0)	26.6 (2.7)	25.9 (3.0)	21.9 (3.9)

Table 2: Demographic details of the LCCN database used to compute the proposed approach. Mean (SD) unless specified otherwise.

	NC	sdMCI	mdMCI	AD
Number of subjects	52	40	58	12
Age, years	69.9 (4.4)	73.8 (6.5)	74.4 (4.3)	75.6 (5.7)
Gender male (%)	15 (29%)	19 (47%)	18 (31%)	7 (58%)
MMSE	29.3 (0.8)	27.4 (2.5)	26.1 (2.7)	22.6 (5.5)

12°; 1 mm slice thickness; a 256×256 matrix; and FOV of 25 cm. The diagnosis of subjects was based on a neuropsychological examination performed at the *Hospital Clínico de Madrid* and the *UPDC del Ayuntamiento de Madrid* [27, 28]. Images are recollected from 52 control subjects, 98 with MCI, and 12 with AD.

The MCI group is divided into two subgroups. ADNI divides patients between amnestic and late MCI [30], whereas in LCCN, the patients with MCI are classified between single and multiple-domain amnestic MCI. Therefore, the criteria are different for separating patients with MCI between both databases. We denote these MCI subgroups using the labels I-MCI and II-MCI in ADNI-HarP, while in LCCN, the labels are sdMCI and mdMCI. Clinical and demographic data are presented in Tables 1 and 2.

On the other hand, the 45 ADNI images of patients with MCI labeled manually using the HarP annotations were selected to be representative of different levels of atrophy in the hippocampus regardless of the type of MCI subgroup that they were classified from ADNI [31]. For this reason, the number of samples in each MCI subgroup in ADNI-HarP is unbalanced.

3. Method

Inspired by work in patch labeling [10, 32], we have recently proposed a new label fusion method for segmenting anatomical structures with low intensity contrast, such as the hippocampus, using T1W-MRI [33, 34]. We developed a patch-based labeling method that cooperates with atlas-warping using non-rigid registrations. The patch-based labeling methods have the advantages of considering multiple samples during the labeling estimation and the local context is well represented by the patches, particularly with affine transformations. In contrast, the label fusion methods using non-rigid registrations lead to segmentations with shape prior constraints. When the delineation of the anatomical structures do not rely on intensity contrast, as in the case of hippocampal segmentation from T1W-MRI, the conventional patch-based labeling is not sufficient for obtaining good results. We have experimentally observed that the collaboration between these two approaches produces higher quality segmentations [33, 34].

3.1. Overview

We first train the segmentation algorithm using the ADNI-HarP images with two manual protocols, which are applied to optimize the various methodological options and parameters. The segmentation accuracy is directly measured by means of the Dice coefficient and volume difference using a leave-one-out analysis. We then evaluate the performance on images belonging to the ADNI-HarP and LCCN databases using two bio-markers to discriminate between patients with AD or MCI and elderly controls.

The proposed segmentation scheme involves four principal steps, namely, (1) MRI pre-processing, (2) spatial normalization and defining the regions of interest (ROIs), (3) the first labeling based on atlas-warping using non-rigid registrations and (4) the final labeling by patches based on similarity measures in intensity and labeling.

During pre-processing of the databases, non-brain regions are removed from all structural images and intracranial volumes are estimated. The images are skull-stripped using BET [35]. Whole brain segmentations are performed in native space. Hippocampal segmentations are performed by spatially normalizing all images to the same stereotactic space [36]. After spatial normalization, a region of interest is defined for each structure studied (left and right hippocampus). For each ROI, the normalized atlas images are first ranked based on their similarity according to the normalized target

image, and the first N_R selected atlases are registered non-rigidly into the ROI of the normalized target image. Next, the registered atlases are fused, and the labeling is calculated using graph cuts based on minimizing an energy function [33]. Then, a patch-based labeling method is applied using the above segmentation of the target image. The patches are selected from the first N_A selected atlases. Fig. 1 shows a flow chart that summarizes the processing of the images.

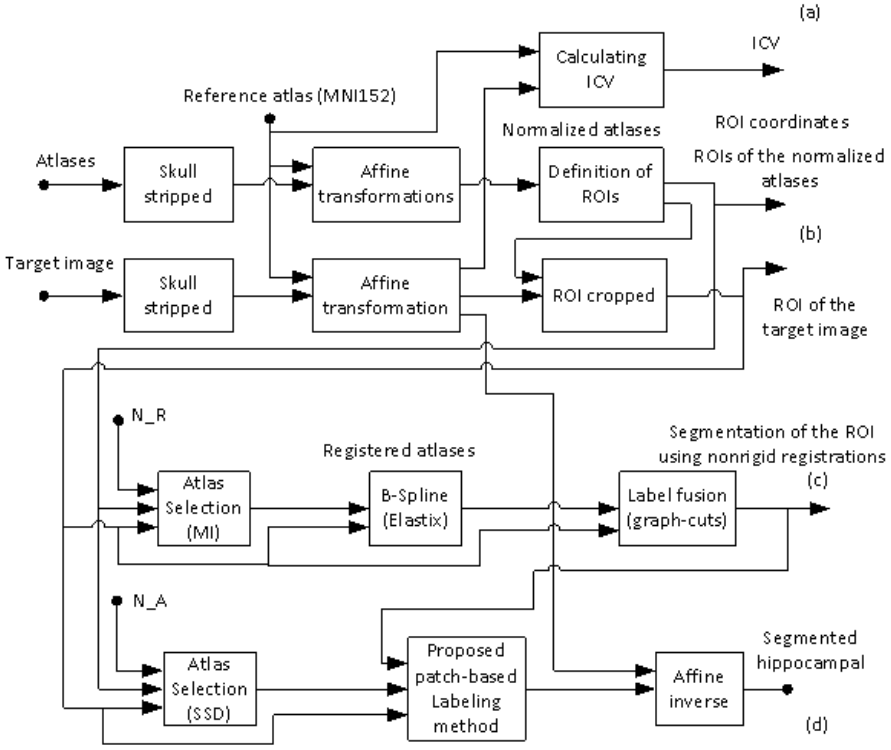

Figure 1: Flow chart summarizing the processing of the patient images: (a) Estimation of the intracranial volume (ICV). (b) MRI pre-processing, spatial normalization and defining the ROIs. (c) Segmenting the ROIs using the non-rigid registration-based label fusion method. (d) Segmenting the ROIs using the proposed patch-based labeling method.

3.2. Skull-stripping

Accurate brain extraction is an important initial step for the success of the following stages. Removing non-brain tissue prior to registration is generally accepted as a means to simplify the inter-subject registration problem and

thus increase the quality of the registrations [37, 38]. Additionally, a key issue with brain extraction tools is their ability to perform adequately when there are varying amounts of cerebral atrophy present, such as that in our case due to AD. Robustness and accuracy of an automated brain extraction method are crucial to reduce manual intervention.

Studies comparing some of the most widely used automated brain extraction techniques show that intensity-based T1 skull-stripping algorithms have the same type of segmentation errors because there is little contrast between cerebrospinal fluid (CSF) and background [39]. Non-brain tissues are included in most automated segmentation algorithms [40]. Some authors [41, 39] developed brain extraction methods that operate on proton density (PD) or T2-weighted images. However, there are two reasons that limit the use of these methods for brain extraction: (i) most MR image processing methods developed thus far are based on the T1-weighted sequence, and (ii) PD and T2-weighted images are not available in all databases.

In our experiments, non-brain regions are removed from all structural images using the open source FSL 4.1 distribution Brain Extraction Tool (BET) [35]. We use BET because it is the best at removing non-brain tissues [40]. BET estimates the minimum and maximum intensity values of the brain image and evolves a deformable model to fit the brain surface based on smoothness criteria and a local intensity threshold. This method starts by finding the center coordinates and tessellates the brain surface using connected triangles [42].

To improve the results of BET, we follow the protocol proposed by Stein et al [43]. First, a coarse skull-stripping is obtained. Then, the bias field correction is calculated and applied to the image. Finally, the corrected image is reapplied with BET (see Fig. 2).

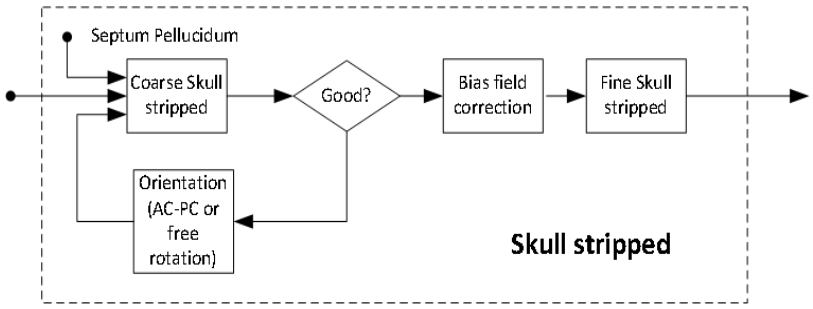

Figure 2: Flow chart summarizing the processing of the skull-stripping.

We have to investigate how to select good center coordinates using the option '-c' and the fractional intensity threshold option '-f' for BET. When applying the first BET to the T1 images, the results are unsatisfactory due to the inclusion of peri-orbital fat, eyes and other non-brain structures in some cases. Some authors use the '-R' option to calculate a robust brain center estimation. We choose to correct such errors by fixing a good center coordinate manually. About the septum pellucidum is chosen as the seed point.

Occasionally, a complete failure of this technique occurs due to poor orientation of the patient's head. We observed that a better orientation of the head improves the results of BET. In these cases, a rotation matrix is generated in 3D-Slicer software through manual placement of landmarks. In the LCCN database, the anterior and posterior commissures are used to align the MRIs along the plane that intersected both points [44]. In the ADNI images, a slight rotation is applied to align the head. These differences in the procedure are due to the LCCN images being less processed than the ADNI images. Atlases and target images are then reoriented using trilinear interpolation by applying the manually generated rotation matrices using FLIRT [45].

BET shows a tendency to exclude some brain voxels in the results. Because the documentation for BET states that a smaller fractional intensity threshold returns a larger brain region, we varied the fractional intensity thresholds between 0.2 and 0.4. We set the '-f' parameter to 0.3 for obtaining the coarse skull-stripping, which matches the tuning of this parameter by Leung et al [40]. These settings were determined by hand-tuning using a subset of subjects and then applied to all subjects, followed by visual inspection of the results.

MR images are generally degraded by a smoothing due to frequency coil non-uniformities that modulate the intensity of the images. Although these artifacts do not typically interfere with visual inspection, they can adversely affect the performance of downstream processing, such as skull stripping and tissue class segmentation [46]. We use FMRIB's Automated Segmentation Tool v.4.1 (FAST), the bias field correction provided in the FSL package [47]. It is well known that FAST is more accurate when the brain has previously been segmented from the background [47]. For this reason, we perform a coarse removal of the brain, and then FAST is applied on this result. After the bias field correction is applied, it proceeds to a second skull stripping. It is a finer extraction with a higher threshold value in the '-f' option, which is

tuned to 0.4.

3.3. The intracranial volume

Normalization by intracranial volume (ICV) reduces the variability in the volume measurements of nearly all brain regions [48]. ICV has been recognized as a suitable constant for normalizing the size of individual brain structures. Compared to other commonly used constants, ICV is less vulnerable to pathological changes [41].

A direct calculation of ICV may be performed from the output of the previous stage, i.e., of the skull stripping. However, using BET on T1-weighted images, we found that a considerable portion of CSF may be excluded, and this constitutes a source of systematic error in the ICV estimation. To overcome these drawbacks, ICV is calculated through registration of each skull-stripped MRI to an atlas-representative template using affine transformations [49]. MNI152 is chosen as the standard brain image [36], and FSL FLIRT is the tool for calculating the affine transformations [45]. The method uses the inverse of the determinant of the transformation matrix, which is multiplied by the template volume ($1.948.105\,mm^3$). The estimation of ICV using this method allows minimization of the bias of the intracranial volume due to the demented older adults with AD or MCI [49]. Furthermore, the spatial normalization of this form is then used for the following steps of the hippocampal segmentation.

3.4. Spatial normalization and atlas registration

Many approaches identify the bio-markers in the brain structures by spatially normalizing all images to the same stereotactic space [50, 51, 52, 53]. To adjust for the differences in head position and size, all images are registered using 12 degrees of freedom brain-to-brain [54]. In our case, the skull-stripped MRIs are affinely registered to the freely downloadable MNI 152 template with a 1 mm^3 isotropic resolution [36] using FLIRT with 12 degrees of freedom [45]. After spatial normalization, a region of interest is defined for each structure studied (left and right hippocampus) as the minimum bounding box containing the structure for all of the training atlases expanded by three voxels along each dimension [53].

For each ROI, the normalized atlas images are first ranked based on their similarity according to the normalized target image using the mutual information (MI) measure [55]. Then, the first N_R atlases are registered non-rigidly into the ROI of the normalized target image. All non-rigid registrations are

computed using *Elastix* [56], a publicly available package for medical image registration. The non-rigid registration of the images is based on the maximization of MI, in combination with a deformation field parameterized by cubic B-splines [57]. The MI is implemented according to [58] using a joint histogram size of 32 x 32 and cubic B-spline Parzen windows. A unique resolution is employed using a B-spline control point spacing of 3.0 mm in all directions. To optimize the cost function, an iterative stochastic gradient descent optimizer is used [59]. In each iteration, 2000 random samples are used to calculate the derivative of the cost function. A maximum of 500 iterations of the stochastic optimization procedure is used. The above-described settings were determined through trial-and-error experiments on two image pairs.

The atlas-labeled images are modeled using the logarithm of odds (LogOdds) formulation, which is based on the signed distance transform [60]. This representation replaces the labels by the signed distances, which are assumed to be positive inside the structure of interest. We found that the LogOdds model produces more accurate results compared with trilinear interpolation or nearest-neighbor interpolation for transferring the atlas-labeled images [61].

3.5. Label propagation

We use a patch-based labeling method, which cooperates with atlas-warping using non-rigid registrations. First, a subset of N_R atlases is registered non-rigidly into the target image, and a label fusion method is applied. In our experiments, the first 15 atlases more similar to the target image were registered non-rigidly ($N_R = 15$) [33]. The label fusion method is based on minimizing a pseudo Boolean function using graph cuts with information of appearance, shape and context [8, 15, 62, 19, 33, 34].

Then, a patch-based labeling method is applied using the above segmentation of the target image. An intensity normalization is applied to the above normalized atlas images using the histogram matching algorithm [63] with the ROI of the normalized target image as the reference. Now, the sum of the squared difference (SSD) measure between each atlas image and the target image is used to rank and select the first N_A atlases. This measure is chosen because SSD is related to the similarity between patches in intensities. The patches of the subset of N_A atlases, which are registered by affine transformations, are pre-selected with a structural similarity measure that takes both the intensity and the labeling of the candidate patches into account [34]. A

compromise between the performance of the patch-based labeling method and computational efficiency is to choose the first 10 atlases that most resemble the target image [64, 65] ($N_A = 10$). From the selected patches, a multi-point label estimation is calculated for each voxel belonging to the target image [64]. The weights of the selected patches are computed from a combination of L^2-norm measures between patches using intensity-based distances and labeling-based distances. Finally, the segmentation result is obtained using the proposed patch-based labeling method, and an inverse affine transformation is applied to return the automatic segmentation into the native space of the target image. For more details about the proposed labeling approach, see [33, 34].

3.6. Classification experiments

Five classification experiments are performed to evaluate and compare the different hippocampal segmentation algorithms. The first experiment is the classification between NC subjects and patients with AD and is referred to as NC vs AD in the following. The next two classifications are between NC and the MCI subgroups. Because the subdivision criteria of patients with MCI are different between the ADNI-HarP and LCCN databases, the comparison groups are not the same. In ADNI-HarP, the classifications are NC vs I-MCI and NC vs II-MCI, whereas in LCCN, the comparisons of patient groups are NC vs single and multiple-domain amnestic MCI. The fourth classification is between the subgroups of patients with MCI, i.e., I-MCI vs II-MCI and sdMCI and mdMCI. The last experiment is between a subgroup of patients with MCI (II-MCI or mdMCI) vs AD.

Our validation framework is designed to compare the capability of different hippocampal segmentation algorithms to discriminate between patients and controls. The bio-markers used are as follows: normalized hippocampal volume and hippocampal surface roughness. Recent studies have shown that a decrease in hippocampal volume and an increase in its surface roughness are good bio-markers for the progress of AD [22]. Moreover, these bio-markers are scalar features, showing robustness in generalization and avoiding overfitting when the size of the samples is limited.

3.6.1. Volume

We first test the classification accuracy using the hippocampal volume. For each subject, we compute the volume of the hippocampus. Volumes are

normalized by the total intracranial volume. For more robustness with respect to segmentation errors, the left and right volumes are averaged. We also compare our approach with the hippocampal volume obtained with the FreeSurfer image analysis suite and corrected with the total intracranial volume also obtained with FreeSurfer.

3.6.2. Shape

We use the surface roughness of the hippocampus as a bio-marker for detecting significant differences between groups of subjects. The hippocampal surface roughness measures its atrophy with the progression of AD using a single scan. The surface roughness is calculated using the mean curvature. The surface roughness of a surface is given by

$$R = \sqrt{\frac{1}{n}\sum_{i}^{n} K^2(x_i)},$$

where n is the number of the voxels belonging to the hippocampal surface and $K(x_i)$ is the mean curvature at each voxel x_i. These voxels are extracted from the automated hippocampal segmentation in the normalized spatial, i.e., $K(x_i)$ is calculated with the isotropic spacing ($1\times1\times1mm^3$ from the MNI 152 space). The left and right hippocampal segmentations are embedded in a level set formulation, and the mean curvature is estimated using:

$$K(x_i) = -div\left(\frac{\nabla\Gamma(x_i)}{\|\nabla\Gamma(x_i)\|}\right),$$

where $\Gamma = \{x|\varphi(x) = 0\}$ is the hippocampal surface and $\varphi(x)$ is a signed distance function, which assigns positive distances to the inside of the object and negative distances to the outside [66]. The estimation of the mean curvature on each x_i is controlled using the Gaussian derivatives. In our experiments, the standard deviation of the Gaussian kernel was set to 2 using trial and error adjustment. For more robustness with respect to segmentation errors, the left and right surface roughnesses are averaged for determining the shape bio-marker.

Fig. 3 shows the average mean curvature over the hippocampal surface of the patient groups. The manual segmentations of ADNI-HarP are used to calculate the mean curvatures of the diagnostic groups. To estimate the average per group, we perform volume-to-volume correspondence [67]. Our non-rigid registration is used for propagating the mean curvature map to

a common template. This activity was performed only for the purpose of displaying the curvature map. It was not used for analysis tasks. As shown in the maps of the weighted curvature of each group, the changes in the values of the mean curvatures are extended in large parts of the hippocampus, particularly in the CA1 and subiculum subfields.

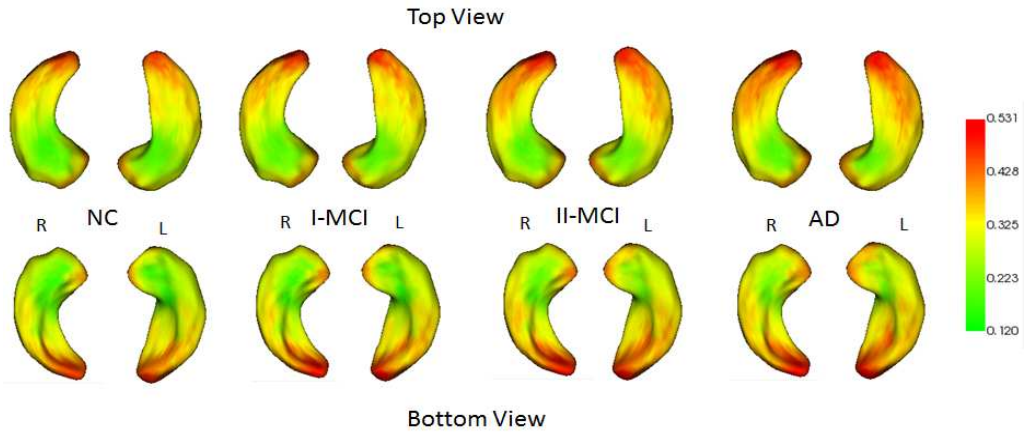

Figure 3: Average mean curvatures for the left and right hippocampus in patients with NC, I-MCI, II-MCI and AD obtained from ADNI-HarP using level set formulation and non-rigid registration into a template.

3.6.3. Evaluation

We use linear discrimination, which fits a normal density to each group, with a pooled estimate of covariance. The use of a simple, linear classifier ensures that the classification accuracy is primarily determined by the quality of the input data rather than by stochastic variations in the classifier. To obtain unbiased estimates of the outcome of the classifiers, leave-one-out cross validation is applied.

For each method and each comparison between diagnostic groups, we computed the number of true positives (TP, i.e., the number of diseased individuals that were correctly identified by the classifier), the number of true negatives (TN, i.e., the number of healthy individuals that were correctly identified by the classifier), the number of false positives (FP, i.e., the number of healthy individuals that were not correctly identified by the classifier) and the number of false negatives (FN, i.e., the number of diseased individuals that were not correctly identified by the classifier). We then

computed the sensitivity, defined as TP/(TP+FN), the specificity, defined as TN/(TN+FP), the positive predictive value, defined as PPV=TP/(TP+FP), the negative predictive value, defined as NPV=TN/(TN+FN), and the accuracy value, defined as ACC=TN+TP/(TN+FP+TP+FN). Finally, note that the number of subjects in each group is not the same. The classification accuracy does not enable the performances between the different classification experiments to be compared. Thus, we considered both the specificity and the sensitivity instead.

4. Results

Two types of experiments are performed using the two methods to compare (a) FreeSurfer v5.3 (FS) and (b) our approach. First, the quality of the automatic segmentations is measured using the DICE coefficient and the relative absolute volume difference [68]. Then, the diagnostic group separation capabilities given by the two methods are evaluated. Furthermore, two different manual segmentations from ADNI images are used to analyze the effect of the manual protocols [12]: (i) 134 labeling of the left and right hippocampus following the HarP [31] (http://www.hippocampal-protocol.net) and (ii) 20 manual segmentations from Automatic Hippocampal Estimator using Atlas-based Delineation (AHEAD) (http://www.ntric.org/projects/ahead).

4.1. Accuracy of the automated segmentations

We compare the automated hippocampal segmentations using the 134 images belonging to ADNI-HarP. For each method in this experiment, we report the DICE coefficient and the relative absolute volume difference as segmentation quality measures [68]. These measures are estimated using leave-one-out cross-validation. Table 3 presents the quantitative segmentation results for each method in the ADNI-HarP images (see Fig. 4 for the distributions of both measures).

Our approach with HarP atlases provides the best results with a mean DICE value of 0.850 and a volume difference of 5.3%. Conversely, FreeSurfer returns average values of 0.782 and 17.8%, respectively. These results are not surprising given that the conventions of manual protocols of the hippocampus differ among the methods. The apparent significant improvement of our method with the HarP, relative to the other two methods, is biased by the choice of the ground-truth segmentations. In this experiment, the HarP has been used as the ground-truth segmentations.

Table 3: Average values and standard deviations of the DICE coefficient and the relative absolute volume difference for all 134 images belonging to the ADNI-HarP database using FreeSurfer and the proposed method with AHEAD atlases and the HarP atlases. The statistical values of the volume difference are calculated as absolute values. However, a signed number is reported in the table so one can recognize under-segmentations by negative values and over-segmentation by positive values.

Status	Dice coefficient			Volume difference (%)		
	FS	AHEAD	HarP	FS	AHEAD	HarP
NC	0.801 ± 0.022	0.729 ± 0.040	0.862 ± 0.017	$+18.9 \pm 10.5$	-32.1 ± 5.9	-4.8 ± 4.0
I-MCI	0.786 ± 0.033	0.729 ± 0.033	0.854 ± 0.020	$+19.1 \pm 8.8$	-30.0 ± 4.1	-4.9 ± 2.8
II-MCI	0.767 ± 0.049	0.734 ± 0.023	0.851 ± 0.033	$+16.0 \pm 9.8$	-29.2 ± 5.5	$+5.6 \pm 3.4$
AD	0.770 ± 0.023	0.725 ± 0.032	0.841 ± 0.024	$+17.6 \pm 9.6$	-27.3 ± 6.6	$+5.9 \pm 5.2$
Global	0.782 ± 0.041	0.730 ± 0.045	0.850 ± 0.024	$+17.8 \pm 9.8$	-29.7 ± 6.1	$+5.3 \pm 4.2$

A qualitative comparison is presented in Fig. 5, which shows the segmentation results of FreeSurfer and of our approach with the HarP for one typical patient for each group (NC, I-MCI, II-MCI and AD). The subject identifications of the ADNI data are as follows: 011_S_0021 (NC), 023_S_0030 (I-MCI), 100_S_0892 (II-MCI) and 003_S_1059 (AD). The first row shows the image that belongs to the ROI in the left hippocampus. In the second row, the image is overlapped with the manual protocol (HarP). The following rows show the segmentations of FreeSurfer (FS) and our approach (PT). The results show that the proposed segmentations globally preserve the hippocampal shape and that the hippocampal surfaces are smoothed relative to the manual segmentations.

4.2. Classification results

Figs. 6 and 7 show the distributions of the normalized hippocampal volume and surface roughness among the diagnostic groups for both databases. Fig. 8 shows that the surface roughness decreased as the volume increases for all patient groups and methods. The results of the classification experiments are summarized in Tables 4, 5, 6 and 7 for both databases and the two bio-markers used.

For each pairwise group to be diagnosed using any of the segmentation methods and both databases, the following guidelines are observed: a) the highest scores are obtained in the pairwise group by comparison to those that are more different. b) The sensibility scores of the normalized hippocampal volume are generally higher than the sensitivity scores of the surface roughness. By contrast, the specificity scores of the surface roughness are generally higher than the specificity scores of the volume bio-marker. Therefore, these

two bio-markers could be complementary to improve the diagnosis. c) The specificity scores are typically higher than the sensitivity scores in the surface roughness bio-marker, and d) the classifiers show bias in their predictive values (i.e., PPV or VPN) when the number of samples between the two comparison groups are very different. This effect is observed in the disparate scores of PPV and NPV.

In general, our approach with the HarP atlases presents slightly better classification results between diagnostic groups than FreeSurfer for both bio-markers used in relation to the classification accuracy. Comparing NC vs AD, the two bio-markers obtained with FreeSurfer and our approach return similar levels of discrimination. By contrast, the proposed method improves the discrimination in NC vs MCI or MCI vs AD in relation to FreeSurfer, with increases of various absolute points in the classification accuracies. Indeed, if the subject groups are more different (i.e., NC vs AD), then the scores are higher and there is less margin for improvement. The scores decrease when comparing NC vs MCI or MCI vs AD. However, our approach generally improves the classification results in comparison to FreeSurfer when the subject groups are more difficult to diagnose. Specifically, using the ADNI-HarP images and normalized hippocampal volume as a bio-marker, the improvements in the classification accuracy in NC vs AD, NC vs II-MCI, NC vs I-MCI, I-MCI vs II-MCI and II-MCI vs AD are 1%, 5%, 7%, 13% and 6%, respectively. In LCCN images with surface roughness as a bio-marker, the improvements in the classification accuracy in the above groups for comparison are 2%, 4%, 3%, 2% and 5%, respectively. Moreover, note that the sensitivity and specificity scores for both bio-markers and both databases are generally higher in our approach compared to FreeSurfer.

Worse results are obtained if the HarP atlases are replaced in our approach by the AHEAD atlases. These results verify that multi-atlas segmentation methods are highly dependent on the manual delineation protocols of the structures to be segmented.

A specific analysis of the results obtained using our method is presented. The classification results for NC vs AD show the highest accuracy with respect to the other comparative groups. In both databases, the bio-marker of surface roughness provides slightly better results than the normalized volume in NC vs AD according to the classification accuracy. The scores provide a high sensibility (over 83%) and specificity (over 87%).

In the comparison between NC and patients MCI with lowest Mini-Mental State Examination (MMSE) values or multiple-domain amnestic (i.e. NC vs

II-MCI in ADNI-HarP or NC vs mdMCI in LCCN), the accuracy scores on both databases are similar, and the volume bio-marker yields results with at least 75% sensibility and 77% specificity.

The classification results are more mixed in the comparison between NC and the subgroup of patients with MCI with higher MMSE scores or single-domain amnestic (i.e., NC vs I-MCI in ADNI-HarP or NC vs sdMCI in LCCN). In the ADNI-HarP images, the normalized volume provides better scores than the surface roughness with 65% sensibility and 70% specificity, while in the LCCN images, both bio-markers return similar accuracy scores although the hippocampal volume gives slight better results than the surface roughness with 66% sensibility and 77% specificity.

The worst scores are obtained in the separation between subgroups of patients with MCI. While there are slight differences between the results of the ADNI-HarP and LCCN images in the other comparative groups, the classification results between subgroups of MCI show some discrepancies. In the ADNI-HarP data and for I-MCI vs II-MCI, the best results are achieved with the normalized volume (67% sensibility and 69% specificity). By contrast, in the LCCN images, the two bio-markers are slightly more accurate than a random classifier. These results show that the criteria for dividing patients with MCI into two subgroups are different in the ADNI-HarP and LCCN databases.

Regarding the comparison between patients MCI with the lowest MMSE values or multiple-domain amnestic vs AD (i.e., II-MCI vs AD in ADNI-HarP or mdMCI vs AD in LCCN), the scores of the classification accuracy are similar for both bio-markers and both databases (65 %). As usual, the best sensitivity scores are with the volume bio-marker (at least 67%), and the best specificity scores are with the surface roughness bio-marker (over 64%).

In summary, for any of the comparisons between subject groups, the classification accuracy results of the two bio-markers are very similar in our approach, particularly with LCCN images. The proposed hippocampal segmentation algorithm estimates the values of the surface roughness and normalized hippocampal volume with a high linear correlation. Fig. 8 shows the relationship between these two markers on ADNI-HarP and LCCN images using our approach and the comparison with FreeSurfer. The linear correlation between these markers in our approach is very high. Whereas FreeSurfer offers greater dispersion between normalized volume and surface roughness, the proposed approach shows less dispersion. The high correlation with the

Table 4: Results of the classifications between patients with different statuses using the normalized hippocampal volume as a bio-marker in the ADNI-HarP database.

Patient classification	Method	SEN	SPE	PPV	NPV	ACC
NC & AD	FS	84%	82%	83%	83%	83%
	AHEAD	83%	75%	72%	82%	79%
	HarP	88%	80%	83%	86%	84%
NC & II-MCI	FS	68%	74%	62%	79%	72%
	AHEAD	75%	75%	66%	83%	75%
	HarP	77%	77%	67%	85%	77%
NC & I-MCI	FS	63%	61%	38%	81%	62%
	AHEAD	58%	56%	36%	76%	56%
	HarP	65%	70%	46%	83%	69%
I-MCI & II-MCI	FS	58%	54%	65%	46%	56%
	AHEAD	66%	59%	73%	54%	65%
	HarP	67%	69%	79%	56%	69%
II-MCI & AD	FS	67%	48%	69%	46%	60%
	AHEAD	68%	54%	70%	52%	63%
	HarP	70%	60%	75%	59%	66%

volume bio-marker indicates that the surface roughness is validated as an efficient bio-marker because its capability to discriminate is similar to that of the hippocampal volume, but it possesses the added advantage of providing a local analysis, which can produce atrophy maps of the progression of AD. Moreover, the proposed estimate of the mean curvature in the image domain is robust, particularly with our segmentation algorithm. We hypothesize that the smoothing of the hippocampal surface provided by our approach leads to the surface roughness being estimated with high reliability (see Fig. 5).

4.3. Computational time

The highest consumption of computational time in our approach occurs in the skull-stripping step. The time required to place the seed in the septum pellucidum and the combined application of BET and FAST consume an average of 10 minutes per patient. The normalization of the brain into the MNI152 template with FLIRT requires less than 3 minutes. The computational complexity of the labeling is primarily due to the non-rigid registrations of the selected atlases into the target image. The computational time for seg-

Table 5: Results of the classifications between patients with different statuses using the normalized hippocampal volume as a bio-marker in the LCCN database.

Patient classification	Method	SEN	SPE	PPV	NPV	ACC
NC & AD	FS	85%	86%	58%	96%	86%
	AHEAD	90%	87%	62%	97%	88%
	HarP	92%	88%	64%	98%	89%
NC & mdMCI	FS	71%	79%	79%	71%	75%
	AHEAD	72%	71%	73%	70%	72%
	HarP	75%	82%	83%	74%	78%
NC & sdMCI	FS	67%	77%	67%	77%	73%
	AHEAD	60%	73%	62%	72%	68%
	HarP	66%	77%	67%	75%	72%
sdMCI & mdMCI	FS	48%	45%	58%	36%	47%
	AHEAD	47%	49%	58%	38%	48%
	HarP	54%	51%	63%	42%	53%
mdMCI & AD	FS	50%	65%	23%	86%	62%
	AHEAD	57%	63%	26%	88%	63%
	HarP	67%	65%	29%	90%	65%

Table 6: Results of the classifications between patients with different statuses using the surface roughness as a bio-marker in the ADNI-HarP database.

Patient classification	Method	SEN	SPE	PPV	NPV	ACC
NC & AD	FS	83%	89%	89%	83%	86%
	AHEAD	79%	81%	81%	79%	80%
	HarP	86%	87%	88%	85%	87%
NC & II-MCI	FS	58%	80%	65%	76%	72%
	AHEAD	65%	76%	64%	78%	72%
	HarP	68%	85%	73%	81%	78%
NC & I-MCI	FS	52%	61%	34%	77%	58%
	AHEAD	54%	60%	34%	77%	57%
	HarP	55%	63%	37%	78%	61%
I-MCI & II-MCI	FS	56%	60%	69%	46%	58%
	AHEAD	60%	59%	71%	49%	60%
	HarP	61%	62%	72%	50%	62%
II-MCI & AD	FS	58%	60%	71%	45%	59%
	AHEAD	55%	54%	66%	43%	55%
	HarP	65%	64%	76%	52%	65%

Table 7: Results of the classifications between patients with different statuses using the surface roughness as a bio-marker in the LCCN database.

Patient classification	Method	SEN	SPE	PPV	NPV	ACC
NC & AD	FS	88%	88%	64%	97%	88%
	AHEAD	74%	86%	55%	93%	84%
	HarP	83%	91%	70%	96%	90%
NC & mdMCI	FS	67%	82%	81%	69%	74%
	AHEAD	72%	75%	76%	71%	74%
	HarP	73%	83%	82%	74%	78%
NC & sdMCI	FS	50%	80%	64%	70%	68%
	AHEAD	60%	73%	62%	72%	68%
	HarP	61%	77%	66%	74%	71%
sdMCI & mdMCI	FS	49%	51%	60%	40%	50%
	AHEAD	45%	48%	57%	36%	46%
	HarP	52%	52%	63%	42%	52%
mdMCI & AD	FS	60%	59%	25%	86%	60%
	AHEAD	49%	65%	23%	86%	62%
	HarP	60%	68%	25%	87%	65%

mentation increases linearly with the number of atlases that have to be registered. However, due to the availability and low cost of multi-core processors, this approach is becoming more feasible. The task of non-rigid registrations has been parallelized. The registration of the first 15 atlases in a ROI requires less than 45 seconds ([Dual CPU] Intel Xeon E5520 @ 2.27 GHz). In the patch-based labeling method, we also parallelize the labeling for each voxel. Our approach takes an average of approximately 4 minutes for fusing labels (including non-rigid registrations) of both the left and right hippocampus on images belonging to ADNI and LCCN. The average time for the hippocampal segmentation from a T1-MRI is less than 17 minutes. The scripts used in this study are available at `https://www.nitrc.org/projects/lf_patches/`.

5. Conclusions

In this paper, we proposed a new hippocampal segmentation algorithm to discriminate patients with AD and MCI from normal aging based on classification of normalized hippocampal volume and hippocampal surface roughness. Based on the training sample of 134 subjects from ADNI-HarP, we developed a new approach for the automatic hippocampal segmentations from T1-MRI. This approach is based on a multiple-atlas segmentation technique, which combines a patch-based labeling method with an atlas-warping using non-rigid registrations. Our method yields mean DICE coefficients of 0.850 and volume differences of approximately 5% with respect to the manual segmentations.

In the two segmentation algorithms (FreeSurfer, our approach with AHEAD atlases and HarP atlases) and the two database used (ADNI and LCCN), we found the lowest hippocampal volume in the AD group and the greatest in the NC group. The average hippocampal volume of patients with MCI was in between the other two groups. We also observed differences across MCI subgroups, with hippocampal volumes lower in those subjects who progressed to a diagnosis of AD compared to those who remained stable (Fig 6). Although the difference in hippocampal volume had the same trend in many previous studies [9, 12, 69], the reported volumes substantially differed. The remarkable difference among studies could be due to the manual segmentation protocols used to define the atlases. However, data from hippocampal volumes are shown to be more homogeneous among the studies when hippocampal volumes are normalized using the intracranial volumes.

Recently, several studies have focused on 3D hippocampal shape analysis. These techniques require the correspondence between the hippocampal segmentation of each subject to a template, and then a local analysis is performed [70, 6, 69]. In contrast, the surface roughness does not require any non-rigid registration. The curvature maps are not propagated to any template, and these maps also offer local information. The surface roughness bio-marker shows an increase in its value with the progression of AD (Fig. 7). Higher values of the surface roughness are in the AD group and the lowest in NC, remaining intermediate values for subgroups of MCI.

The proposed hippocampal segmentation algorithm estimates the values of the surface roughness and normalized hippocampal volume with high linear correlation (Fig. 8). However, the classification results show that these bio-markers are complementary. The hippocampal surface roughness generally offers higher specificity scores than the normalized hippocampal volume; conversely, the sensitivity score of the hippocampal volume is typically higher than that produced by the surface roughness. We consider that both the procedure for estimating the mean curvature and the smoothing of the hippocampal surface given by our approach lead to the surface roughness being estimated with high reliability (see Fig. 5). Furthermore, our surface roughness has the ability to diagnose AD similar to the hippocampal volume, but it has the added advantage of providing a local analysis, which can produce atrophy maps of the progression of AD.

As shown in maps of the weighted curvature of each group, the atrophy patterns are extended in large parts of the hippocampus, particularly in the CA1 and subiculum subfields (Fig. 3). These subfields are typically reported as the earliest and most significant atrophic regions in AD [20, 71, 69]. Therefore, the surface roughness is a very robust scalar measure that finely quantifies the atrophy caused by AD. Statistical analysis for searching significance maps between diagnosed groups are left for future work.

For our approach with HarP annotations, for all compared groups and the two image databases used, the classification results with both bio-markers provide slightly better results than those obtained with FreeSurfer. The classification improvements of our approach are more substantial when subject groups are more difficult to diagnose. The impact in our method of replacing the atlas set using HarP by another set of atlases (AHEAD) is also shown. The classification results are worse when using AHEAD atlases. Definitively, multi-atlas segmentation methods are highly dependent on manual segmentation protocols. Our approach achieved classification accuracies of 90% for

AD vs NC, 78% for NC vs MCI with the lowest MMSE values or multiple-domain amnestic, 70% for NC vs MCI with the highest MMSE values or single-domain amnestic and 65% for AD vs MCI with the lowest MMSE values or multiple-domain amnestic. The scores are similar in both databases for these compared groups. The worst results are obtained in the separation between subgroups of MCI, and there are score discrepancies between the two databases. While the classification accuracy is of 69% between subgroups of patients with MCI in the ADNI-HarP data, in the LCCN images, the two bio-markers are slightly more accurate than a random classifier, i.e., these bio-markers are not able to discriminate between patients with single and multiple-domain amnestic MCI. The MMSE scores for MCI subgroups indicate that the criteria used to generate the subgroups are different in the ADNI-HarP and LCCN databases (see Tables 1 and 2).

To further validate the developed classification models, the ADNI-HarP database could have been used as the training set and the LCCN images used as a test set (and vice versa) to observe how well the models can predict new and unseen data. However, the differences in criteria for labeling patients into subgroups of MCI did not allow the two databases to be joined. However, the comparison between NC vs AD is feasible, and the classification results were similar to those obtained by cross validation. Regardless, the similarity of the classification results in both databases, bio-markers and segmentation methods employed shows unbiased scores largely because the two bio-markers used are scalars, showing robustness in generalization and avoiding overfitting.

We conclude that our approach with HarP annotations has a high level of accuracy for the segmentation of the hippocampus and is robust to multi-site data. Our automatically obtained regions can be used to obtain several bio-markers of the hippocampus from T1-MRI. Our tool will enable fast analysis of imaging data for AD progression. The average time for the hippocampal segmentation from a T1-MRI is less than 17 minutes. We also left for future work the application of our approach into longitudinal studies. The scripts used in this study are available at
`https://www.nitrc.org/projects/lf_patches/`.

References

[1] Jack, C., Shiung, M., Gunter, J., Obrien, P., Weigand, S., Knopman, D., Boeve, B., Ivnik, R., Smith, G., Cha, R., et al.: Comparison of different

MRI brain atrophy rate measures with clinical disease progression in AD. Neurology **62** (2004) 591–600

[2] Dubois, B., Feldman, H.H., Jacova, C., DeKosky, S.T., Barberger-Gateau, P., Cummings, J., Delacourte, A., Galasko, D., Gauthier, S., Jicha, G., et al.: Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria. The Lancet Neurology **6** (2007) 734–746

[3] Frisoni, G.B., Fox, N.C., Jack, C.R., Scheltens, P., Thompson, P.M.: The clinical use of structural MRI in Alzheimer disease. Nature Reviews Neurology **6** (2010) 67–77

[4] Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.O., Chupin, M., Benali, H., Colliot, O., et al.: Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. Neuroimage **56** (2011) 766–781

[5] Clerx, L., van Rossum, I.A., Burns, L., Knol, D.L., Scheltens, P., Verhey, F., Aalten, P., Lapuerta, P., van de Pol, L., van Schijndel, R., et al.: Measurements of medial temporal lobe atrophy for prediction of Alzheimer's disease in subjects with mild cognitive impairment. Neurobiology of Aging **34** (2013) 2003–2013

[6] Gerardin, E., Chételat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.S., Niethammer, M., Dubois, B., Lehéricy, S., Garnero, L., et al.: Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. Neuroimage **47** (2009) 1476–1486

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

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

[9] Leung, K.K., Barnes, J., Ridgway, G.R., Bartlett, J.W., Clarkson, M.J., Macdonald, K., Schuff, N., Fox, N.C., Ourselin, S., et al.: Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. Neuroimage **51** (2010) 1345–1359

[10] Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L.: Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. Neuroimage **54** (2011) 940–954

[11] Wang, H., Suh, J.W., Das, S.R., Pluta, J.B., Craige, C., Yushkevich, P.A.: Multi-atlas segmentation with joint label fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence **35** (2013) 611–623

[12] Nestor, S.M., Gibson, E., Gao, F.Q., Kiss, A., Black, S.E.: A direct morphometric comparison of five labeling protocols for multi-atlas driven automatic segmentation of the hippocampus in Alzheimer's disease. Neuroimage **66** (2013) 50–70

[13] Konrad, C., Ukas, T., Nebel, C., Arolt, V., Toga, A.W., Narr, K.: Defining the human hippocampus in cerebral magnetic resonance imagesan overview of current segmentation protocols. Neuroimage **47** (2009) 1185–1195

[14] Fischl, B., Salat, D., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S.: Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron **33** (2002) 341–355

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

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

[17] Boccardi, M., Ganzola, R., Bocchetta, M., Pievani, M., Redolfi, A., Bartzokis, G., Camicioli, R., Csernansky, J.G., de Leon, M.J., de Toledo-Morrell, L., et al.: Survey of protocols for the manual segmentation of the hippocampus: preparatory steps towards a joint EADC-ADNI harmonized protocol. Journal of Alzheimer's Disease **26** (2011) 61–75

[18] Frisoni, G.B., Jack, C.R., Bocchetta, M., Bauer, C., Frederiksen, K.S., Liu, Y., Preboske, G., Swihart, T., Blair, M., Cavedo, E., et al.: The EADC-ADNI Harmonized Protocol for manual hippocampal segmentation on magnetic resonance: Evidence of validity. Alzheimer's & Dementia **11** (2015) 111–125

[19] Wolz, R., Heckemann, R.A., Aljabar, P., Hajnal, J.V., Hammers, A., Lötjönen, J., Rueckert, D.: Measurement of hippocampal atrophy using 4D graph-cut segmentation: application to ADNI. Neuroimage **52** (2010) 109–118

[20] Csernansky, J., Wang, L., Swank, J., Miller, J., Gado, M., McKeel, D., Miller, M., Morris, J.: Preclinical detection of Alzheimer's disease: hippocampal shape and volume predict dementia onset in the elderly. Neuroimage **25** (2005) 783–792

[21] Gutman, B., Wang, Y., Morra, J., Toga, A.W., Thompson, P.M.: Disease classification with hippocampal shape invariants. Hippocampus **19** (2009) 572–578

[22] Kim, J., Valdes-Hernandez, M., Royle, N., Park, J.: Hippocampal shape modeling based on a progressive template surface deformation and its verification. IEEE Transactions on Medical Imaging **34** (2015) 1242–1261

[23] Van Leemput, K., Bakkour, A., Benner, T., Wiggins, G., Wald, L.L., Augustinack, J., Dickerson, B.C., Golland, P., Fischl, B.: Automated segmentation of hippocampal subfields from ultra-high resolution in vivo MRI. Hippocampus **19** (2009) 549–557

[24] Iglesias, J.E., Augustinack, J.C., Nguyen, K., Player, C.M., Player, A., Wright, M., Roy, N., Frosch, M.P., McKee, A.C., Wald, L.L., et al.: A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: Application to adaptive segmentation of in vivo MRI. Neuroimage **115** (2015) 117–137

[25] Yushkevich, P.A., Wang, H., Pluta, J., Das, S.R., Craige, C., Avants, B.B., Weiner, M.W., Mueller, S.: Nearly automatic segmentation of hippocampal subfields in vivo focal T2-weighted MRI. Neuroimage **53** (2010) 1208–1224

[26] Yushkevich, P.A., Pluta, J.B., Wang, H., Xie, L., Ding, S.L., Gertje, E.C., Mancuso, L., Kliot, D., Das, S.R., Wolk, D.A.: Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment. Human Brain Mapping **36** (2015) 258–287

[27] Pineda-Pardo, J.A., Garcés, P., López, M.E., Aurtenetxe, S., Cuesta, P., Marcos, A., Montejo, P., Yus, M., Hernández-Tamames, J.A., del Pozo, F., et al.: White matter damage disorganizes brain functional networks in amnestic mild cognitive impairment. Brain connectivity **4** (2014) 312–322

[28] Pineda-Pardo, J.A., Bruña, R., Woolrich, M., Marcos, A., Nobre, A.C., Maestú, F., Vidaurre, D.: Guiding functional connectivity estimation by structural connectivity in meg: an application to discrimination of conditions of mild cognitive impairment. Neuroimage **101** (2014) 765–777

[29] Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L Whitwell, J., Ward, C., et al.: The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. Journal of Magnetic Resonance Imaging **27** (2008) 685–691

[30] Aisen, P.S., Petersen, R.C., Donohue, M.C., Gamst, A., Raman, R., Thomas, R.G., Walter, S., Trojanowski, J.Q., Shaw, L.M., Beckett, L.A., et al.: Clinical core of the Alzheimer's Disease Neuroimaging Initiative: progress and plans. Alzheimer's & Dementia **6** (2010) 239–246

[31] Boccardi, M., Bocchetta, M., Morency, F.C., Collins, D.L., Nishikawa, M., Ganzola, R., Grothe, M.J., Wolf, D., Redolfi, A., Pievani, M., et al.: Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol. Alzheimer's & Dementia **11** (2015) 175–183

[32] Coupé, P., Eskildsen, S.F., Manjón, J.V., Fonov, V.S., Collins, D.L., disease Neuroimaging Initiative, A., et al.: Simultaneous segmentation and grading of anatomical structures for patient's classification: application to Alzheimer's disease. Neuroimage **59** (2012) 3736–3747

[33] Platero, C., Tobar, M.C.: A label fusion method using conditional random fields with higher-order potentials: Application to hippocampal segmentation. Artificial Intelligence in Medicine **64** (2015) 117–129

[34] Platero, C., Tobar, M.C.: Combining patch based strategies and non-rigid registration-based label fusion methods. ArXiv e-prints (2015)

[35] Smith, S.M.: Fast robust automated brain extraction. Human Brain Mapping **17** (2002) 143–155

[36] Evans, A.C., Janke, A.L., Collins, D.L., Baillet, S.: Brain templates and atlases. Neuroimage **62** (2012) 911–922

[37] Battaglini, M., Smith, S.M., Brogi, S., De Stefano, N.: Enhanced brain extraction improves the accuracy of brain atrophy estimation. Neuroimage **40** (2008) 583–589

[38] Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M.C., Christensen, G.E., Collins, D.L., Gee, J., Hellier, P., et al.: Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. Neuroimage **46** (2009) 786–802

[39] Ramirez, J., Gibson, E., Quddus, A., Lobaugh, N.J., Feinstein, A., Levine, B., Scott, C., Levy-Cooperman, N., Gao, F., Black, S.E.: Lesion explorer: a comprehensive segmentation and parcellation package to obtain regional volumetrics for subcortical hyperintensities and intracranial tissue. Neuroimage **54** (2011) 963–973

[40] Leung, K.K., Barnes, J., Modat, M., Ridgway, G.R., Bartlett, J.W., Fox, N.C., et al.: Brain MAPS: an automated, accurate and robust brain extraction technique using a template library. Neuroimage **55** (2011) 1091–1108

[41] Pengas, G., Pereira, J., Williams, G.B., Nestor, P.J.: Comparative reliability of total intracranial volume estimation methods and the influence of atrophy in a longitudinal semantic dementia cohort. Journal of Neuroimaging **19** (2009) 37–46

[42] Clark, K.A., Woods, R.P., Rottenberg, D.A., Toga, A.W., Mazziotta, J.C.: Impact of acquisition protocols and processing streams on tissue segmentation of T1 weighted MR images. Neuroimage **29** (2006) 185–202

[43] Stein, J.L., Medland, S.E., Vasquez, A.A., Hibar, D.P., Senstad, R.E., Winkler, A.M., Toro, R., Appel, K., Bartecek, R., Bergmann, Ø., et al.: Identification of common variants associated with human hippocampal and intracranial volumes. Nature Genetics **44** (2012) 552–561

[44] Talairach, J., Tournoux, P.: Co-planar stereotaxic atlas of the human brain. 3-dimensional proportional system: an approach to cerebral imaging. (1988)

[45] Jenkinson, M., Bannister, P., Brady, M., Smith, S.: Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage **17** (2002) 825–841

[46] Cohen, M.S., DuBois, R.M., Zeineh, M.M.: Rapid and effective correction of RF inhomogeneity for high field magnetic resonance imaging. Human Brain Mapping **10** (2000) 204–211

[47] Zhang, Y., Brady, M., Smith, S.: Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Transactions on Medical Imaging **20** (2001) 45–57

[48] Keihaninejad, S., Heckemann, R.A., Fagiolo, G., Symms, M.R., Hajnal, J.V., Hammers, A., Initiative, A.D.N., et al.: A robust method to estimate the intracranial volume across MRI field strengths (1.5T and 3T). Neuroimage **50** (2010) 1427–1437

[49] Buckner, R.L., Head, D., Parker, J., Fotenos, A.F., Marcus, D., Morris, J.C., Snyder, A.Z.: A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume. Neuroimage **23** (2004) 724–738

[50] Ashburner, J., Friston, K.J.: Voxel-based morphometrythe methods. Neuroimage **11** (2000) 805–821

[51] Bozzali, M., Filippi, M., Magnani, G., Cercignani, M., Franceschi, M., Schiatti, E., Castiglioni, S., Mossini, R., Falautano, M., Scotti, G., et al.: The contribution of voxel-based morphometry in staging patients with mild cognitive impairment. Neurology **67** (2006) 453–460

[52] Tu, Z., Narr, K., Dollár, P., Dinov, I., Thompson, P., Toga, A.: Brain anatomical structure segmentation by hybrid discriminative/generative models. IEEE Transactions on Medical Imaging **27** (2008) 495–508

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

[54] Woods, R.P., Grafton, S.T., Holmes, C.J., Cherry, S.R., Mazziotta, J.C.: Automated image registration: I. General methods and intrasubject, intramodality validation. Journal of Computer Assisted Tomography **22** (1998) 139–152

[55] Viola, P., Wells III, W.M.: Alignment by maximization of mutual information. International Journal of Computer Vision **24** (1997) 137–154

[56] Klein, S., Staring, M., Murphy, K., Viergever, M., Pluim, J.: Elastix: a toolbox for intensity-based medical image registration. IEEE Transactions on Medical Imaging **29** (2010) 196–205

[57] Rueckert, D., Sonoda, L., Hayes, C., Hill, D., Leach, M., Hawkes, D.: Nonrigid registration using free-form deformations: application to breast MR images. IEEE Transactions on Medical Imaging **18** (1999) 712–721

[58] Thévenaz, P., Unser, M.: Optimization of mutual information for multiresolution image registration. IEEE Transactions on Image Processing **9** (2000) 2083–2099

[59] Klein, S., Staring, M., Pluim, J.: Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines. IEEE Transactions on Image Processing **16** (2007) 2879–2890

[60] Pohl, K.M., Fisher, J., Shenton, M., McCarley, R.W., Grimson, W.E.L., Kikinis, R., Wells, W.M.: Logarithm odds maps for shape representation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI. Springer (2006) 955–963

[61] Sabuncu, M., Yeo, B., Van Leemput, K., Fischl, B., Golland, P.: A generative model for image segmentation based on label fusion. IEEE Transactions on Medical Imaging **29** (2010) 1714–1729

[62] Song, Z., Tustison, N., Avants, B., Gee, J.: Integrated graph cuts for brain MRI segmentation. Medical Image Computing and Computer-Assisted Intervention–MICCAI **4191** (2006) 831–838

[63] Gonzales, R.C., Woods, R.E.: Digital image processing. New Jersey: Prentice Hall **6** (2002) 1–689

[64] Rousseau, F., Habas, P.A., Studholme, C.: A supervised patch-based approach for human brain labeling. IEEE Transactions on Medical Imaging **30** (2011) 1852–1862

[65] Tong, T., Wolz, R., Coupé, P., Hajnal, J.V., Rueckert, D.: Segmentation of MR images via discriminative dictionary learning and sparse coding: application to hippocampus labeling. Neuroimage **76** (2013) 11–23

[66] Osher, S., Fedkiw, R.: Level set methods and dynamic implicit surfaces. Volume 153. Springer Science & Business Media (2006)

[67] Heimann, T., Meinzer, H.P.: Statistical shape models for 3D medical image segmentation: a review. Medical image analysis **13** (2009) 543–563

[68] van Ginneken, B., Heimann, T., Styner, M.: 3D segmentation in the clinic: A grand challenge. In: MICCAI Workshop on 3D Segmentation in the Clinic: A Grand Challenge. (2007) 7–15

[69] Frankó, E., Joly, O.: Evaluating Alzheimer's disease progression using rate of regional hippocampal atrophy. PloS one **8** (2013) e71354

[70] Styner, M., Oguz, I., Xu, S., Brechbühler, C., Pantazis, D., Levitt, J.J., Shenton, M.E., Gerig, G.: Framework for the statistical shape analysis of brain structures using SPHARM-PDM. The insight journal (2006) 242

[71] Chételat, G., Fouquet, M., Kalpouzos, G., Denghien, I., De La Sayette, V., Viader, F., Mézenge, F., Landeau, B., Baron, J.C., Eustache, F.,

et al.: Three-dimensional surface mapping of hippocampal atrophy progression from MCI to AD and over normal aging as assessed using voxel-based morphometry. Neuropsychologia **46** (2008) 1721–1731

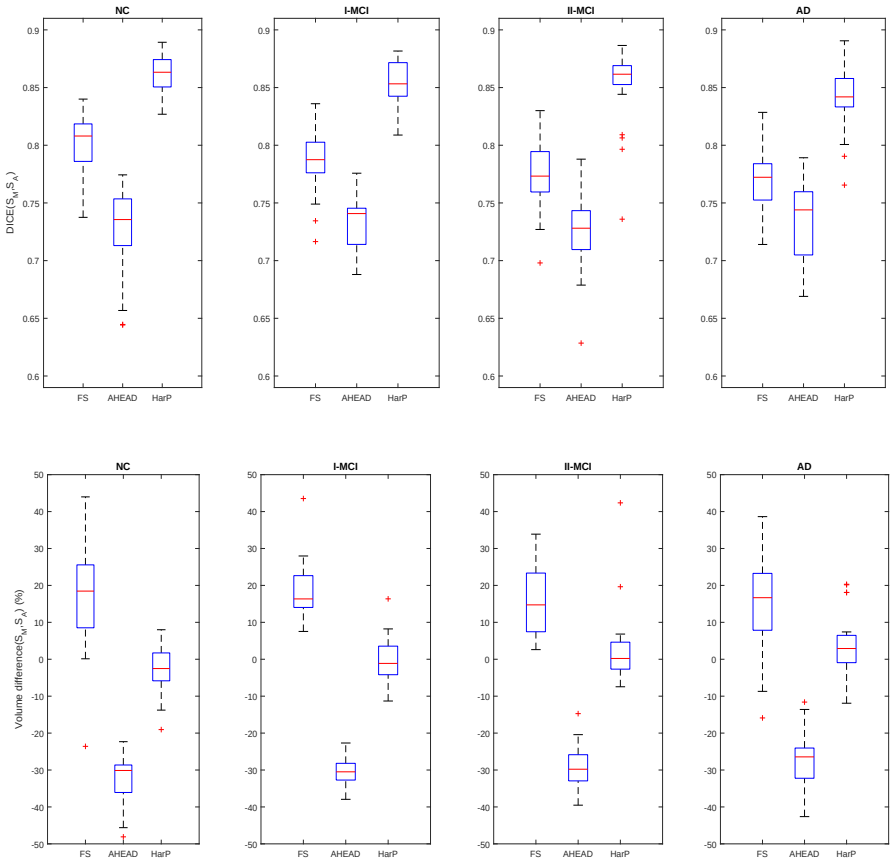

Figure 4: The DICE distribution for 134 images belonging to the ADNI-HarP database using FreeSurfer and the proposed method with AHEAD atlases and HarP atlases. The first row shows the DICE (S_M,S_A) distributions, where S_M is the manual hippocampal segmentation and S_A is the automated hippocampal segmentation. The second row shows the relative absolute volume difference distributions.

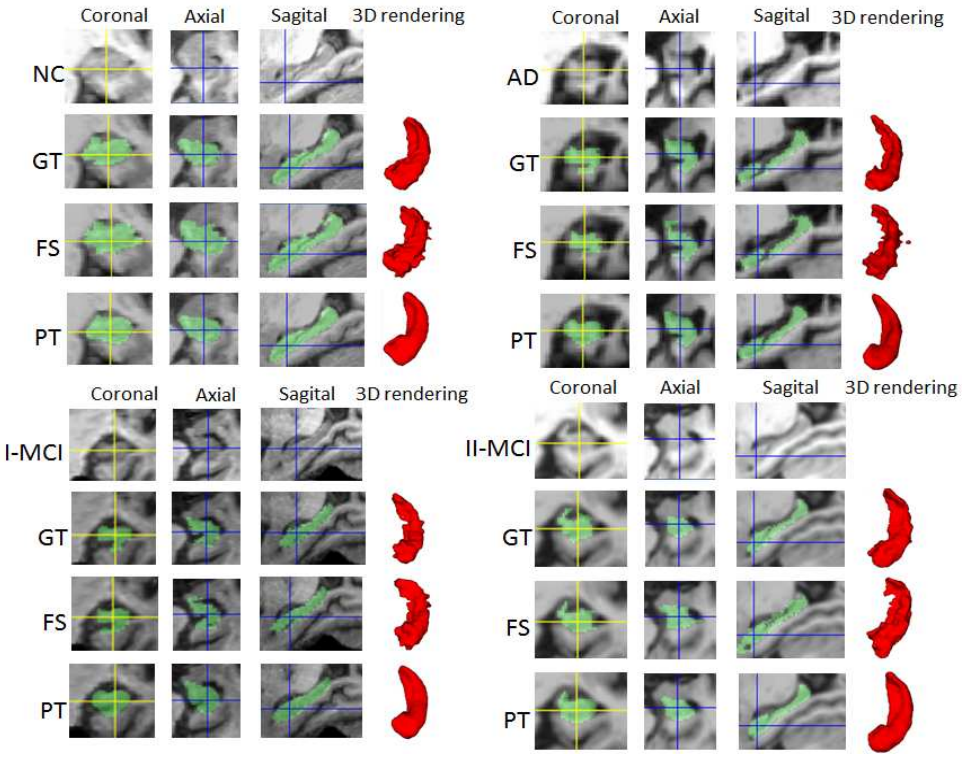

Figure 5: Comparison between FreeSurfer and the proposed method. The first row shows the image belonging to the ROI in the left hippocampus. In the second row, the image is overlapped with the manual protocol (HarP). This is considered the Ground Truth (GT). The following rows show the segmentations of the FreeSurfer (FS) and our approach (PT).

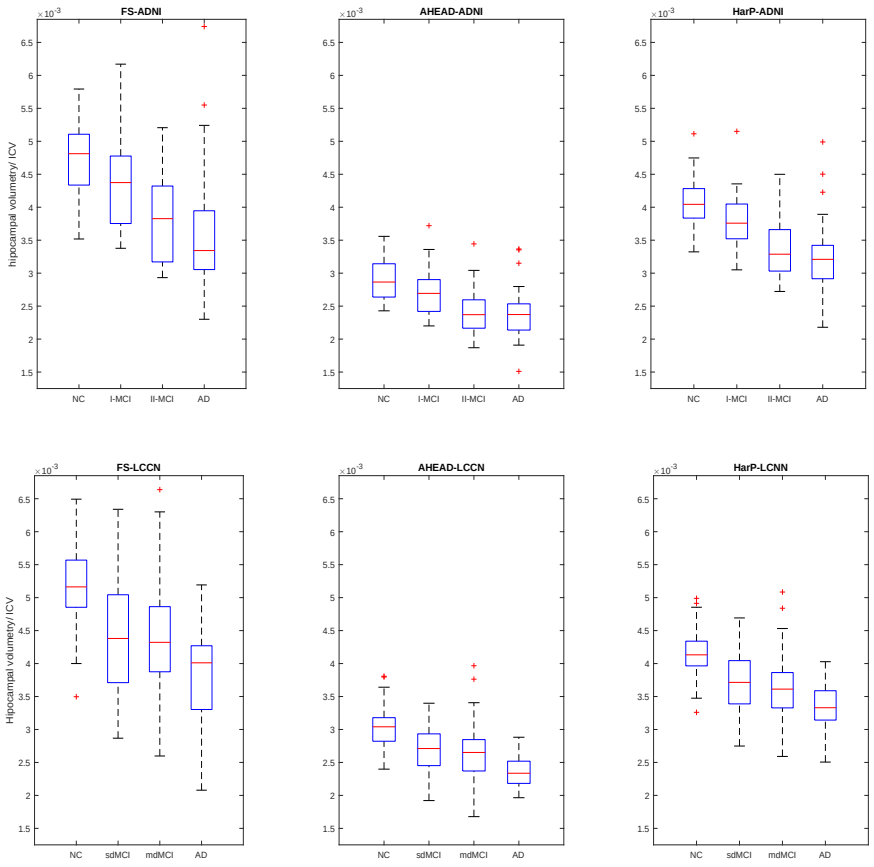

Figure 6: The normalized hippocampal volume distribution for 134 and 162 images belonging to the ADNI and LCCN databases using FreeSurfer and the proposed method with AHEAD atlases and the HarP atlases.

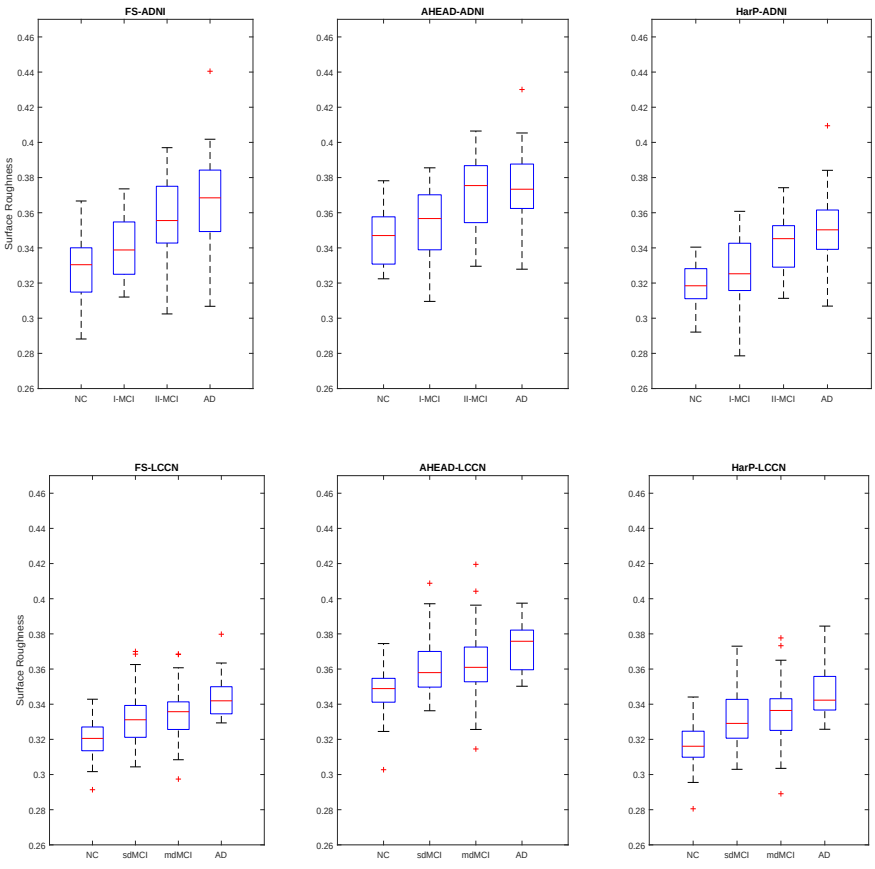

Figure 7: The surface roughness distribution for 134 and 162 images belonging to the ADNI and LCCN databases using FreeSurfer and the proposed method with AHEAD atlases and the HarP atlases.

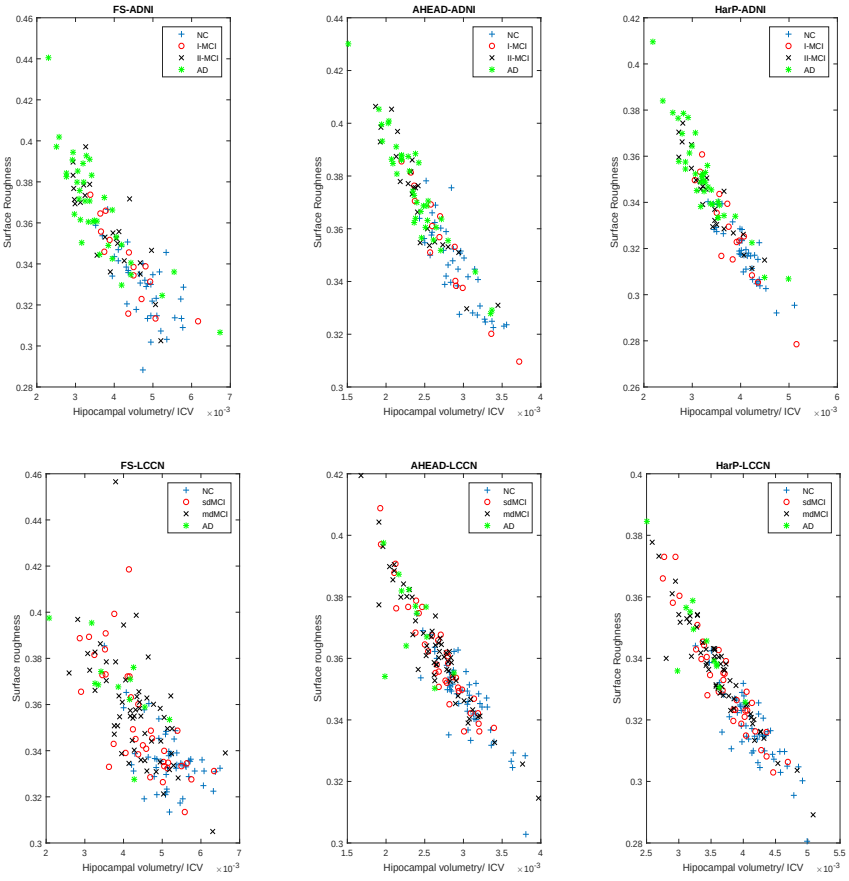

Figure 8: Relation between volume and surface roughness of the automated hippocampal segmentations from the ADNI-HarP and LCCN databases.