Physica Medica 31 (2015) 1085–1091

Contents lists available at ScienceDirect

Physica Medica

journal homepage: http://www.physicamedica.com

Technical Notes

Multiple RF classifier for the hippocampus segmentation: Method and validation on EADC-ADNI Harmonized Hippocampal Protocol

P. Inglese [a,b], N. Amoroso [a,b], M. Boccardi [c], M. Bocchetta [c,d], S. Bruno [e], A. Chincarini [f], R. Errico [b,f,g], G.B. Frisoni [c,h,i], R. Maglietta [j], A. Redolfi [c], F. Sensi [f], S. Tangaro [a,*], A. Tateo [a,b], R. Bellotti [a,b] for the Alzheimer's Disease Neuroimaging Initiative [1]

[a] Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Italy
[b] Università degli Studi di Bari, Bari, Italy
[c] LENITEM Laboratory of Epidemiology, Neuroimaging & Telemedicine, IRCSS Istituto Centro San Giovanni di Dio - Fatebenefratelli, Brescia, Italy
[d] Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy
[e] Overdale Hospital, St Helier, Jersey, UK
[f] Istituto Nazionale di Fisica Nucleare, Sezione di Genova, Italy
[g] Università degli Studi di Genova, Genova, Italy
[h] aFaR Associazione FateBeneFratelli per la Ricerca, Rome, Italy
[i] Psychogeriatric Ward, IRCSS S. Giovanni di Dio – FBF, Brescia, Italy
[j] Istituto di Studi sui Sistemi Intelligenti per l'Automazione, Consiglio Nazionale delle Ricerche, Bari, Italy

ARTICLE INFO

Article history:
Received 25 February 2015
Received in revised form 16 July 2015
Accepted 12 August 2015
Available online 21 October 2015

Keywords:
Hippocampus segmentation
Random forest classifier
Alzheimer's disease

ABSTRACT

The hippocampus has a key role in a number of neurodegenerative diseases, such as Alzheimer's Disease. Here we present a novel method for the automated segmentation of the hippocampus from structural magnetic resonance images (MRI), based on a combination of multiple classifiers. The method is validated on a cohort of 50 T1 MRI scans, comprehending healthy control, mild cognitive impairment, and Alzheimer's Disease subjects. The preliminary release of the EADC-ADNI Harmonized Protocol training labels is used as gold standard. The fully automated pipeline consists of a registration using an affine transformation, the extraction of a local bounding box, and the classification of each voxel in two classes (background and hippocampus). The classification is performed slice-by-slice along each of the three orthogonal directions of the 3D-MRI using a Random Forest (RF) classifier, followed by a fusion of the three full segmentations. Dice coefficients obtained by multiple RF (0.87 ± 0.03) are larger than those obtained by a single monolithic RF applied to the entire bounding box, and are comparable to state-of-the-art. A test on an external cohort of 50 T1 MRI scans shows that the presented method is robust and reliable. Additionally, a comparison of local changes in the morphology of the hippocampi between the three subject groups is performed. Our work showed that a multiple classification approach can be implemented for the segmentation for the measurement of volume and shape changes of the hippocampus with diagnostic purposes.

Introduction

In the last 20 years, the hippocampus has acquired a key role as a biomarker for many neuropsychiatric diseases such as Alzheimer's disease (AD) [1], and major depression. Specifically, in the case of AD, hippocampal shape and volume changes represent as an early indicator of tissue degeneration. Intercepting these changes is of crucial importance, since the timely introduction of disease-modifying treatments may bring to a reduction of progression rate of the illness with improvement of patients' quality of life. To date, manual segmentation performed by trained experts is still considered the reference standard for the hippocampus identification but, although several protocols have been proposed [2], none has emerged as a standard thus far, making it difficult to compare experimental results from different investigators. Trying to fill this gap, EADC-ADNI has developed a new protocol based on the most influential protocols adopted last years. The Harmonized Hippocampus Protocol (HarP) [3] aims to delineate a set of rules for the manual

[*] Corresponding author. Dipartimento Interateneo di Fisica "M. Merlin", Università degli Studi di Bari, Via Amendola 173, 70126 Bari, Italy. Tel.: +39 080 5442370
E-mail address: sonia.tangaro@ba.infn.it (S. Tangaro).

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

http://dx.doi.org/10.1016/j.ejmp.2015.08.003
1120-1797/© 2015 Associazione Italiana di Fisica Medica. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

segmentation of the hippocampus that could be shared among the different research teams. However, manual segmentation can be extremely time consuming, and is therefore impractical for large-scale studies. In light of this, in the last decade, a great effort has been devoted to the development of automatic segmentation techniques showing good predictive performance and reasonable computational times. Among automatic segmentation methods, notable results have been achieved by pattern recognition models, deformable shapes, and multi-atlas label fusion.

In machine-learning domain, frequently real problems involve building large complex models. In these cases, one of the suggested approaches is defining a set of simple models that can easily catch the local properties of data. Multiple classification, which is based on this idea, has been widely used for automatic learning and pattern recognition tasks, showing an improvement of detection performance over single monolithic classifiers. As an example, we may cite widely adopted methods, like "ensembles" (e.g. Random Forest, AdaBoost), are based on simple base learners that perform the same task on different subsets of samples or features. A general idea is that the base learner should provide good performance and a sufficient level of diversity [4], so that coincident errors can be reduced.

In our work, a modular approach was adopted defining three sets of complementary base learners, each of them specialized in segmenting each slice of the 3D MRI scan along its three principal directions. Thereafter, the three full 3D segmentations were combined using a majority voting approach in order to avoid complexities that could reduce the generality of the method [5].

Several examples of applications of multiple classifiers to medical imaging have already been reported in the literature. An approach based on fusion of best performing classifiers to detect breast lesions from Dynamic Contrast-Enhanced MRI (DCE-MRI) scans was proposed in [6]. Other methods are based on fusion of different/complementary feature set classifiers [7,8]. However, the literature lacks examples of multiple classifiers applied to the brain parcelization. The method presented here is similar to [9], where a set of Adaboost classifiers fused with a majority voting rule is applied to human organ localization in 3D Computed Tomography (CT) images. The distinctive feature of our work lies in the extraction of the shape of the sub-cortical region of interest, performing a segmentation of the hippocampal region.

Materials and methods

All 3D MRI scans were obtained from the ADNI database (https://ida.loni.usc.edu), together with the manual segmentations from the preliminary release [10] of EADC-ADNI Harmonized Protocol (HarP) training labels, available at http://www.hippocampal-protocol.net/SOPs/labels.php. HarP has been developed by the major international experts of hippocampal tracing in AD, with the aim of harmonizing the available protocols for manual tracing of the hippocampus in order to create a standard shared protocol.

Two subject cohorts were used in our experiments. To tune the classifiers' parameters and evaluate their performance, we used the first cohort (D1), consisting of 50 3D MRI scans (14 normal control (NC), 17 mild cognitive impairment (MCI), and 19 Alzheimer's Disease (AD) subjects) with age in the range of 60–89 years. A second independent cohort (D2) consisting of 50 3D MRI scans (15 NC, 17 MCI, 18 AD) with age in the range of 61–90 was only used to test the classifiers' performance. D2 images were not used during the training phase.

MRI scans alignment and ROI extraction

The MRI scans were aligned through an affine (12 degrees of freedom) transformation, using the Insight Segmentation and Registration Toolkit (http://www.itk.org/). A shape based pre-segmentation of the hippocampus was then performed (FAPoD) [11] to identify a smaller region of interest encased by a bounding box ($50 \times 60 \times 60$ voxels). The latter procedure aimed to decrease the data dimensionality and consequently the computational time.

Feature extraction

After MRI scans alignment, a set of about 300 features associated to each voxel was computed [12,13]. These consisted of: intensity, gradients, co-occurrence based Haralick features computed along the directions 0°, 45°, 90°, 135°, 1D, 2D, and 3D Haar-like features, mean filters, and variance filters. All the features were computed using a neighborhood kernel centered in the voxel and with size varying from $3 \times 3 \times 3$ to $9 \times 9 \times 9$ voxels.

Automated hippocampal segmentation

In this section we give a description of the proposed multiple classifier method for the automatic segmentation of the hippocampus. RF [14] was selected as base learner for its robustness to noise and overfitting. Furthermore, manual feature selection was avoided since RF was able to perform it intrinsically. The 3D ROI was first split into 1-voxel thick slices along the three orthogonal directions x, y, and z corresponding to the axial, coronal, and sagittal directions, respectively. In this way, each group of orthogonal slices was processed independently.

From the training images, we defined the training voxels for the RF classifier as a mask-filtered subset of each slice. Specifically, we took the area of the slice containing the hippocampus and dilated it by a $\delta \times \delta$ voxel sized square kernel, inclusive of background voxels in the neighborhood of the hippocampus boundaries. In addition, each RF classifier required tuning of two parameters: number t of trees and minimum number l of observation per leaf. Successively, the classifiers were applied to the test image slices along the corresponding direction k, on which a different filtering mask was applied. This was defined as the union of hippocampal regions in the training slices. Since MRI scans were aligned during the registration, we could assume that this region included the hippocampal region of test images.

The final segmentation of the hippocampus was obtained applying a majority voting rule to the union of the three binary segmentations (one for each direction). A graphical scheme of the algorithm as it worked along the direction x is shown in Fig. 1.

Evaluation metrics

The accuracy of the hippocampus automatic segmentation was evaluated by measuring a set of standard metrics. Error, precision, recall, and Dice's similarity coefficient (DSC) are defined as follows:

$$\text{Error} = \frac{\text{FP}+\text{FN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{3}$$

$$\text{DSC} = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}}. \tag{4}$$

where true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are the confusion matrix elements.

Each of these metrics is sensitive to one or more components of the confusion matrix, hence classifier performance should be evaluated jointly with the metrics outcomes. In particular, error rate

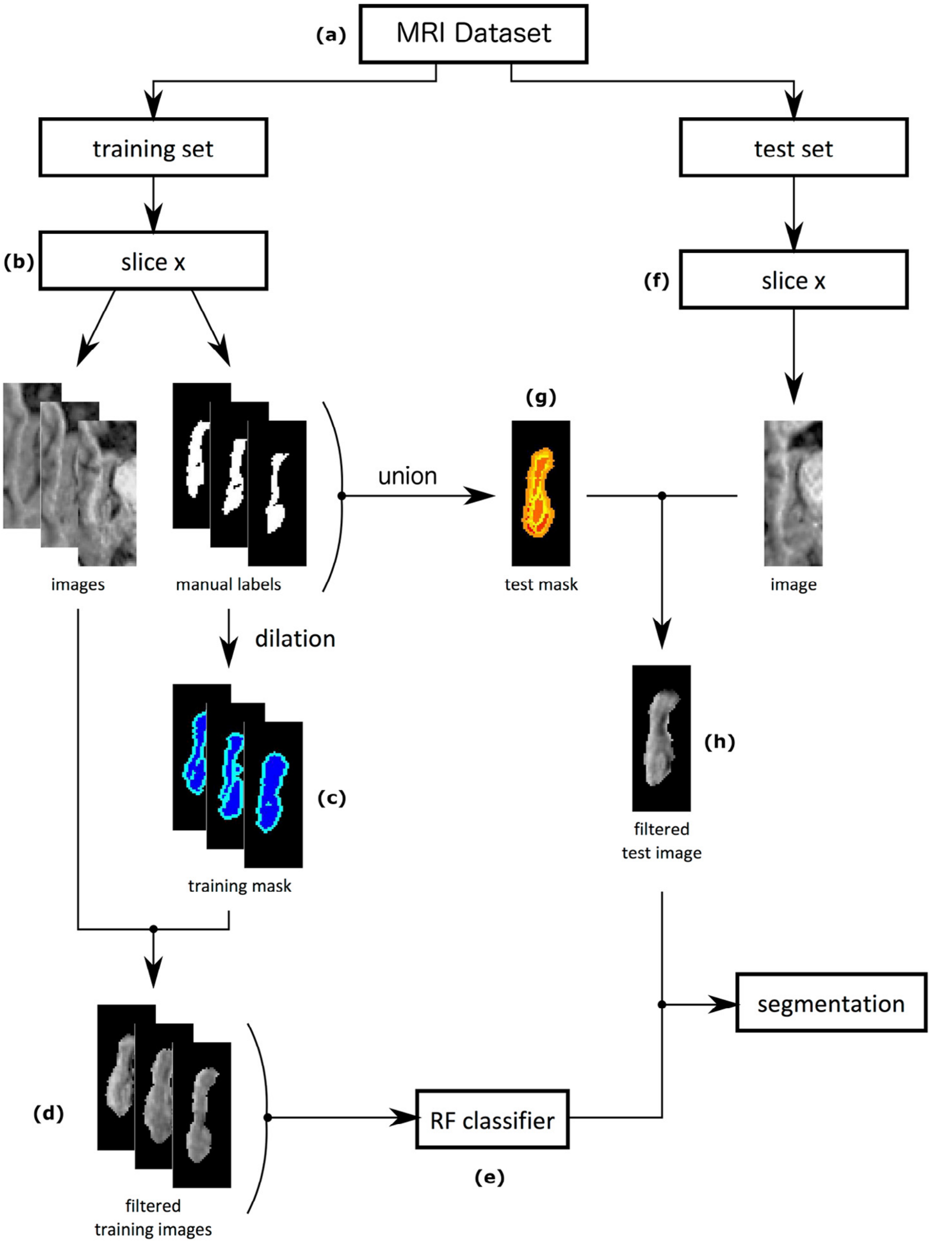

Figure 1. A diagram of the segmentation procedure for the MRI slices taken along the *x* direction. The MRI scans dataset (a) is split into two sets of images corresponding to the training and test set, respectively. All the planes along the *x* axis are separated (b)(f) and each plane number is classified separately. A filtering mask obtained dilating the hippocampal mask (c) is applied to all the training slices (d) which then are used to build an RF classifier (e). The dilated union of all the training hippocampal masks (g) is used to filter the test slices (h). The procedure is repeated for the other two directions and thereafter the three 3D segmentations are fused with a majority vote procedure.

may be underestimated when the test set is very imbalanced. DSC represents a measure of the agreement between automated and manual segmentation. Precision and recall are more sensitive to FP and FN, respectively, and provide a measure of the most frequently mislabeled class.

Local atrophy mapping

The local shape differences between the hippocampi of the three groups were analyzed through the SPHARM-MAT toolbox for Matlab[1]. This method allows, through an implicit description of 3D objects, a study of morphological correspondences. First, binary hippocampal segmentation was converted to a parameterized surface mesh with a 'Topology fix' (connectivity = (6 +, 18), epsilon = 1.5). Thereafter, hippocampal shapes were aligned establishing a correspondence among all the surfaces using a first-order ellipsoid (FOE) algorithm. To assess local differences among the surfaces of the three subject groups, a vertex-by-vertex t-test was performed on the surface manifold and the significance maps of T-values were computed.

Results

Performance evaluation

The model validation was performed with a 10-fold cross validation, using the metrics described in *Evaluation metrics*. Overall

[1] http://imaging.indyrad.iupui.edu/projects/SPHARM/

performance was assessed by averaging those metrics. Furthermore, a number of trees t of 150 and a minimum number of samples per leaf l of 5 gave the minimum out-of-bag error.

Training set properties

Mean hippocampal percentage in the ROI was (3.1 ± 0.1)%, therefore we expected classifiers to be biased toward the majority class, resulting in a poor detection capability. As a measure of the classifier's capability to correctly distinguish between classes, we used *accuracy on positive examples* (equivalent to recall) and *accuracy on negative examples*, also called *specificity*. In particular, we measured the *G-mean* score, defined as the *geometric mean* of the *recall* (see Eq. 3) and *specificity*, which results to being poor when a classifier is biased toward one class,

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{5}$$

$$\text{G-mean} = \sqrt{\text{Recall} \times \text{Specificity}} \tag{6}$$

Effects of imbalance on classifier performance

The effect of class imbalance on classifier performance was evaluated by changing the parameter δ of the filtering mask on the training slices from 1 to 7. In this way, training sets with different class ratios were considered. The filtering mask for the test set was dilated of 2 voxels to take into account the morphological variability among the hippocampi in the dataset.

The error rate and the DSC for different values of the training filter dilation parameter (Fig. 2) showed that the use of a small number of background voxels ($\delta = 4$) in the neighborhood region of the hippocampus can improve the segmentation accuracy.

We also compared the performance of the multiple RF with that of a single RF. In this case, the filtering mask was defined as the hippocampal region dilated by the same δ used in the classification slice-by-slice. The results in Table 1, where only metrics with $\delta = 4$ are reported, show that the performance of multiple RFs was better than a single monolithic classifier. It is also evident (Table 1) that the proposed method performed better than the classifiers built slice-by-slice along a single direction, confirming that the fusion of the three classifiers made it capable of correcting misclassified voxels to each other. The results achieved with our method are in line with the state-of-the-art, showing a DSC of (0.87 ± 0.03) for the left hemisphere and (0.87 ± 0.04) for the right hemisphere. Moreover, by this method we were able to improve the performance obtained in our previous work [15–17], where we used a single RF classifier with an analogous filtering approach. However, in those studies, a different cohort was used and, in particular, all the scans were manually segmented with different protocol. Therefore, this comparison can be considered indicative of an improvement due to the combination of the proposed method and the HarP protocol.

Table 1
Average metrics and standard deviations for left and right hemispheres of D1 (top) ($\delta = 4$) for the multiple classifier (RF_{multi}), the single classifier (RF_{single}), and the classifier built slice-by-slice along a single direction. Average metrics and standard deviations for D2 (bottom).

Dataset D1				
Left hemisphere				
Method	Error	DSC	Precision	Recall
RF_{multi}	0.008 ± 0.002	**0.87 ± 0.03**	0.90 ± 0.04	0.85 ± 0.05
RF_{single}	0.008 ± 0.002	0.86 ± 0.03	0.87 ± 0.05	0.86 ± 0.04
RF_x	0.009 ± 0.002	0.86 ± 0.03	0.87 ± 0.04	0.85 ± 0.04
RF_y	0.008 ± 0.002	0.86 ± 0.03	0.88 ± 0.04	0.85 ± 0.05
RF_z	0.008 ± 0.002	0.86 ± 0.03	0.89 ± 0.04	0.84 ± 0.05
Right hemisphere				
Method	Error	DSC	Precision	Recall
RF_{multi}	0.008 ± 0.002	**0.87 ± 0.04**	0.90 ± 0.05	0.84 ± 0.06
RF_{single}	0.009 ± 0.002	0.85 ± 0.04	0.86 ± 0.06	0.85 ± 0.06
RF_x	0.009 ± 0.002	0.85 ± 0.04	0.87 ± 0.06	0.84 ± 0.06
RF_y	0.009 ± 0.002	0.85 ± 0.04	0.88 ± 0.05	0.84 ± 0.07
RF_z	0.008 ± 0.002	0.86 ± 0.04	0.90 ± 0.05	0.83 ± 0.07
Dataset D2				
Hemisphere	Error	DSC	Precision	Recall
Left	0.008 ± 0.003	0.86 ± 0.04	0.89 ± 0.04	0.85 ± 0.06
Right	0.009 ± 0.004	0.86 ± 0.09	0.88 ± 0.07	0.84 ± 0.11

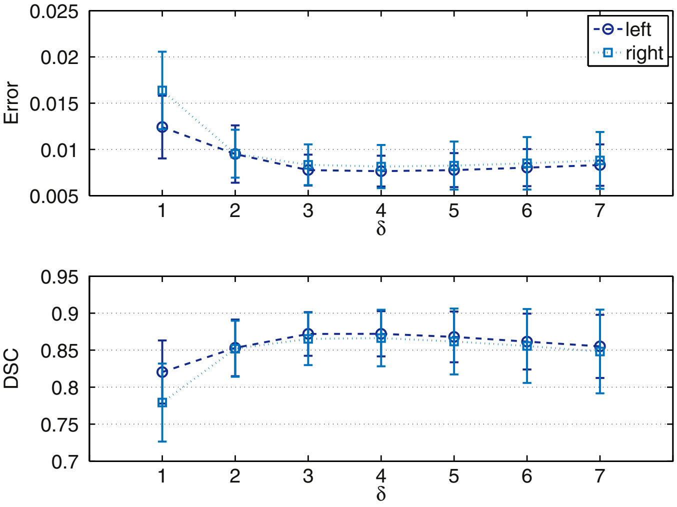

Figure 2. Error (upper panel) and DSC (lower panel) trends varying the values of δ for the left and right hemispheres. The best performance is achieved with a value of $\delta = 4$.

As shown by the effects of the variation of the value of the dilation parameter on the *G-mean* (Table 2), classifiers were more prone to either negative or positive class depending on the class balance in the training set. It is worth noting that the most balanced classifier was obtained with $\delta = 3$, followed by $\delta = 4$, which corresponded to the best DSC. Also, it should be noticed that as the value of δ increases, despite the error and DSC curves showing a decrease in the performance, they do not vary much, confirming the redundancy of the information contained into the most external voxels. Applying FreeSurfer to the dataset we obtained a DSC of 0.74 ± 0.05 for the left hemisphere and 0.76 ± 0.05 for the right hemisphere, confirming that the presented method was able to outperform it.

Volumetric measure

An additional aim of our study was the detection of significant differences of hippocampal volumes between NCs, MCIs, and ADs. The hippocampal volume was obtained by counting the number of voxels classified as hippocampus. Significant differences ($p < 0.05$)

Table 2
Average accuracies on positive (a^+) and negative (a^-) examples, and G-mean score, and standard deviations for different values of δ. Only left hemisphere is reported here.

δ	*G-mean*	a^+	$\boldsymbol{a^-}$
1	0.89 ± 0.03	**0.92 ± 0.03**	0.85 ± 0.05
2	0.91 ± 0.02	0.89 ± 0.04	0.93 ± 0.04
3	**0.92 ± 0.02**	0.86 ± 0.04	0.97 ± 0.01
4	0.91 ± 0.03	0.85 ± 0.05	
5	0.91 ± 0.03	0.83 ± 0.05	0.99 ± 0.01
6	0.90 ± 0.03	0.82 ± 0.06	0.99 ± 0.01
7	0.89 ± 0.04	0.80 ± 0.07	**0.99 ± 0.01**

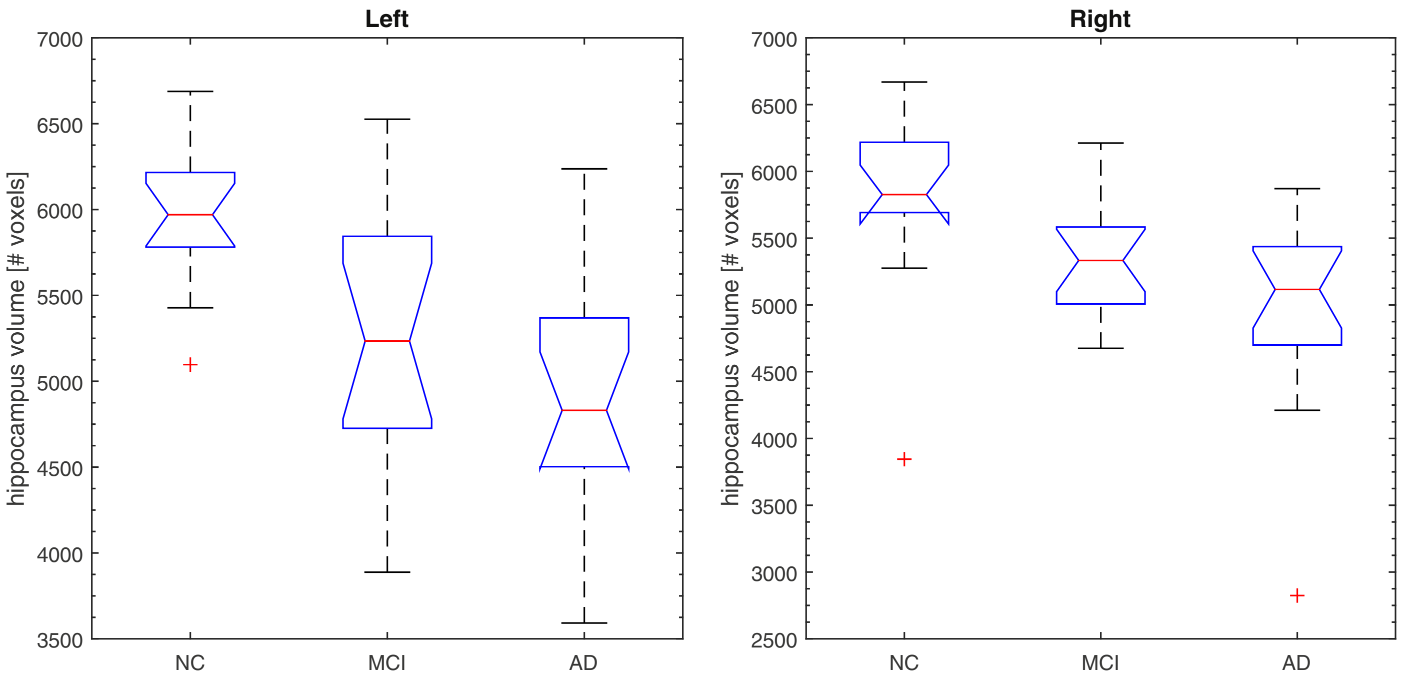

Figure 3. Box plot of the predicted volumes for the three subject groups.

among volumes of NC and MCI, and NC and AD subjects were confirmed with the paired analysis of variance using a one-way ANOVA (Fig. 3), in which p-values, corrected with Tukey–Kramer method, are reported in Table 3. However MCI volumes were not statistically different from those of ADs in the left hemisphere, whereas only the NCs and ADs were statistically different in the right hemisphere.

A comparison of the hippocampal shapes between the three groups was then performed through the SPHARM-MAT toolbox. In Fig. 4, significant T-values ($p < 0.05$) from a paired t-test between the positions of each vertex of the hippocampal surfaces of the three groups are reported.

Test on D2 dataset

To assess the robustness of the presented method, the models built on the cohort D1 were applied on the external cohort D2. All the ROIs were subject to the same preprocessing steps applied to the D1 scans, and the final segmentation was defined by the majority voting among the 10 classifiers built with the 10-fold cross validation on D1. As shown in Table 1, although the performance on D2 was slightly lower than that on D1, as one could expect, it remained comparable. This confirmed the reliability of the entire learning process made on D1.

Table 3
Multiple comparisons between volumes of the three groups of subjects. In the second column the 95% confidence interval of the difference between the volumes of the two groups is shown, and in the third column the estimated mean volume differences are reported. In the fourth column the corrected p-values are reported.

Group difference	95% confidence interval*	Estimated mean	p-value
Left hemisphere			
NC/MCI	[111.4, 1322.0]	716.7	0.02
NC/AD	[485.4, 1677.7]	1081.5	$2*10^{-4}$
MCI/AD	[–220.6, 950.3]	364.9	0.294
Right hemisphere			
Group difference	95% confidence interval*	Estimated mean	p-value
NC/MCI	[–120.1, 1040.7]	460.3	0.143
NC/AD	[276.4, 1419.5]	848.0	0.002
MCI/AD	[–173.6, 949.0]	387.7	0.223

* Lower and upper limits for 95% confidence intervals for the true mean difference.

Conclusion

Hippocampal atrophy is an established biomarker for a number of neurodegenerative pathologies, such as the AD. Here we have proposed an algorithm based on the use of multiple RF classifiers, aimed to segment local portions of the hippocampus. Because of the complex morphological structure of subcortical regions, we have proposed a fusion approach, where the ROI is "seen" by the classifiers from different directions, in order to exploit local shape patterns. To reduce the class imbalance in the ROI, we introduced a filtering mask, modeled on the true hippocampal shapes. A study of the dilation parameter associated to this filter showed that only the voxels in the neighborhood of the hippocampus boundaries are necessary to achieve the best prediction accuracy and the most balanced classifier.

The effectiveness of our method was confirmed by comparing its performance with those of a single RF applied to the entire ROI, subject to the same filtering mask. Indeed, we could observe that the multiple classifiers performed better in both hemispheres. This can be justified by the fact that the multiple classifiers were designed to catch local patterns in contrast to a general model of the entire hippocampus as learned by the single classifier.

Using two datasets consisting of 50 subjects each divided into cohorts of Normal Controls, Mild Cognitive Impairment, and Alzheimer's Disease patients, we assessed the validity of our approach with a 10-fold cross validation. We obtained a DSC of 0.87 for the left and right hemispheres, results that are comparable to state-of-the-art [18], and represented an improvement on our previous work [15–17], where we used a single RF classifier with an analogous filtering approach. The classifier, when used on the second dataset, which was not involved during the training and tuning phase,

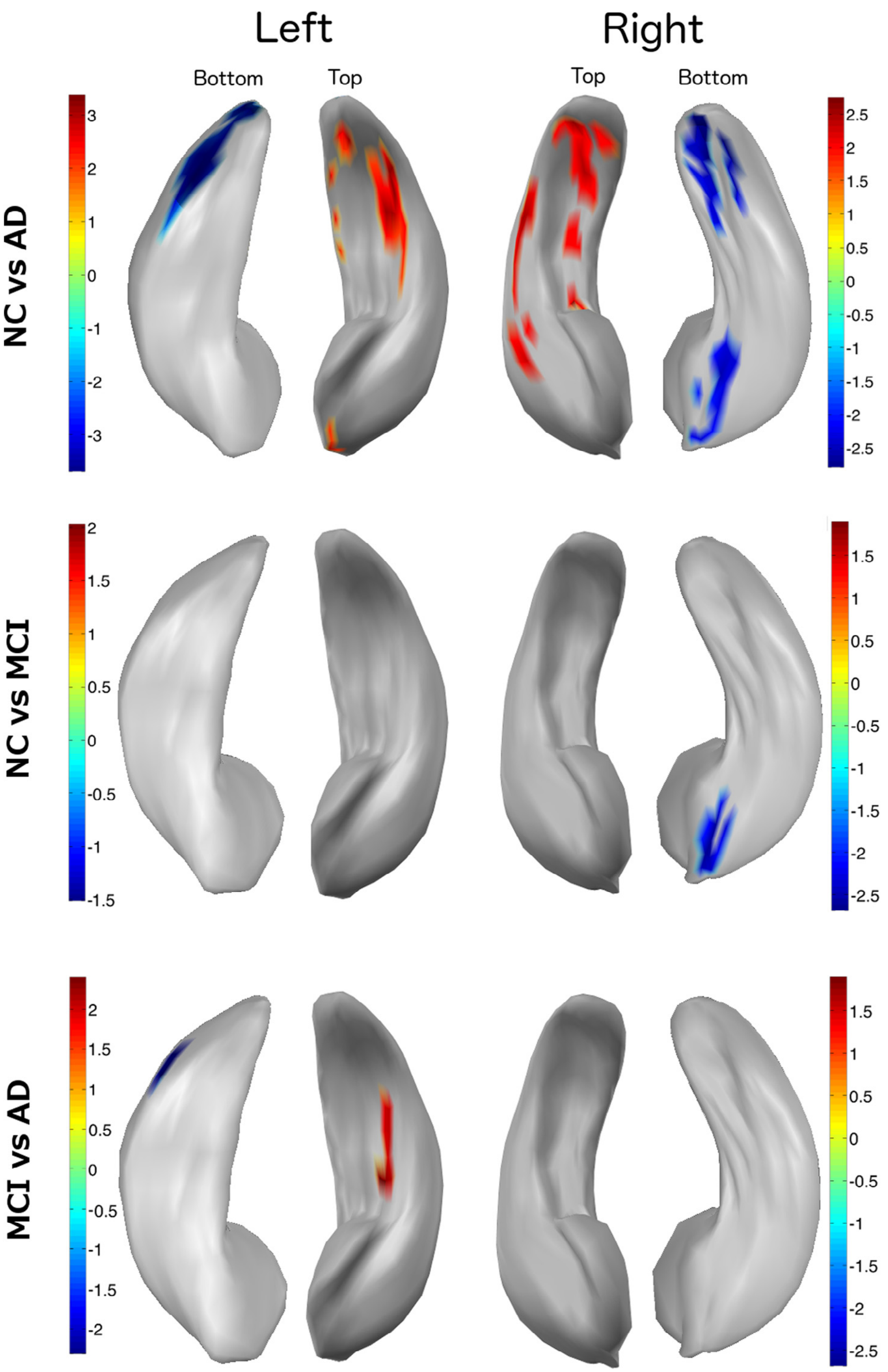

Figure 4. T-value color map is reported. Positive and negative T-values represent the outward and inward shape differences between the three groups of subjects.

showed comparable performance, with a slight, but expected, decreased DSC. Furthermore, the presented method was able to outperform FreeSurfer, which resulted in a DSC of 0.74 ± 0.05 for the left hemisphere and 0.76 ± 0.05 for the right hemisphere.

We also mapped the local differences between the shapes of the three groups' hippocampi. This analysis was intended sorely to confirm the reliability of the proposed method, since we used the raw p-values without multiple test corrections. However, it is interesting to notice that the deformations of the hippocampi obtained by the automated segmentation are consistent with those of [19,20], where NCs and ADs are significantly different, whereas other paired tests do not show significant differences. Furthermore, the local atrophy of the bottom part of the tail and head of the hippocampus was consistent with those previous works.

Another advantage of the presented method relies on its scalability on large datasets, being a slice-by-slice classification followed by a fusion. Indeed, in a cluster environment, we were able to perform the entire classification in the time necessary for the classification of a single slice (about 5 minutes per cross validation round). On the other hand, the same task was performed by the single classifier in about 40 minutes. In view of the EADC-ADNI initiative in developing the HarP protocol as a shared segmentation protocol for the hippocampus, acceleration of the processing time for large datasets can represent an important practical aspect of our work.

Acknowledgment

This research was supported by Istituto Nazionale di Fisica Nucleare (INFN), Italy and by grants from Università degli Studi di Bari, Italy. All authors disclose any actual or potential conflicts of interest, including any financial, personal, or other relationships with other people or organizations that could inappropriately influence their work. All experiments were performed with the informed consent of each participant or caregiver in line with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Local institutional ethics committees approved the study.

Data collection and sharing for this project was funded by ADNI (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from private sector contributions, which are facilitated by the Foundation for the National Institutes of Health.

References

[1] Bron E, Smits M, van der Flier W, Vrenken H, Barkhof F, Scheltens P, et al. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. Neuroimage 2015; 111:562–79.

[2] Boccardi M, Ganzola R, Bocchetta M, Pievani M, Redolfi A, Bartzokis G, et al. Survey of protocols for the manual segmentation of the hippocampus: preparatory steps towards a joint EADC-ADNI Harmonized Protocol. J Alzheimers Dis 2011;26:61–75.

[3] Boccardi M, Bocchetta M, Apostolova LG, Barnes J, Bartzokis G, Corbetta G, et al. Delphi definition of the EADC-ADNI Harmonized Protocol for hippocampal segmentation on magnetic resonance. Alzheimers Dement 2015;11(2):126–38.

[4] Melville P, Mooney RJ. Creating diversity in ensembles using artificial data. Inform Fusion 2005;6(1):99–111.

[5] Ruta D, Gabrys B. A theoretical analysis of the limits of majority voting errors for multiple classifier systems. Pattern Anal Appl 2002;5(4):333–50.

[6] Fusco R, Sansone M, Petrillo A, Sansone C. A Multiple Classifier System for Classification of Breast Lesions Using Dynamic and Morphological Features in DCE-MRI. Proceedings of the 2012 Joint IAPR International Conference on Structural, Syntactic, and Statistical Pattern Recognition, SSPR'12/SPR'12, (Berlin, Heidelberg), Springer-Verlag; 2012. p. 684–92.

[7] Li B, Li W, Zhao D. Multi-scale feature based medical image classification, In Computer Science and Network Technology (ICCSNT), 2013 3rd International Conference on, Oct 2013. p. 1182–6.

[8] Han X-H, Chen Y-W. Biomedical imaging modality classification using combined visual features and textual terms. Int J Biomed Imaging 2011;2011:241396.

[9] Zhou X, Wang S, Chen H, Hara T, Yokoyama R, Kanematsu M, et al. Automatic localization of solid organs on 3D CT images by a collaborative majority voting decision based on ensemble learning. Comput Med Imaging Graph 2012; 36(4):304–13.

[10] Boccardi M, Bocchetta M, Morency FC, Collins DL, Nishikawa M, Ganzola R, et al. Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol. Alzheimers Dement 2015;11(2):183–91.

[11] Amoroso N, Bellotti R, Bruno S, Chincarini A, Logroscino G, Tangaro S, et al. Automated Shape Analysis landmarks detection for medical image processing. Proceedings of the International Symposium, CompIMAGE; 2012. p. 139–42.

[12] Tangaro S, Amoroso N, Brescia M, Cavuoti S, Chincarini A, Errico R, et al. Feature selection based on machine learning in MRIs for hippocampal segmentation. Comput Math Methods Med 2015;2015: http://dx.doi.org/10.1155/2015/814104.

[13] Maglietta R, Amoroso N, Boccardi M, Bruno S, Chincarini A, Frisoni GB, et al. Automated hippocampal segmentation in 3D MRI using random undersampling with boosting algorithm. Pattern Anal Applic 2015. doi: 10.1007/s10044-015-0492-0.

[14] Breiman L. Random forests. Mach Learn 2001;5–32.

[15] Tangaro S, Amoroso N, Boccardi M, Bruno S, Chincarini A, Ferraro G, et al. Automated voxel-by-voxel tissue classification for hippocampal segmentation: methods and validation. Phys Med 2014;30:878–87.

[16] Tangaro S, Amoroso N, Bruno S, Chincarini A, Frisoni G, Maglietta R, et al. Active Learning Machines for Automatic Segmentation of Hippocampus in MRI. In: Industrial Conference on Data Mining – Workshops'13; 2013. p. 181–91.

[17] Maglietta R, Amoroso N, Bruno S, Chincarini A, Frisoni G, Inglese P, et al. Random Forest Classification for Hippocampal Segmentation in 3D MR Images. In: Machine Learning and Applications (ICMLA), 2013 12th International Conference on, vol. 1, Dec 2013. p. 264–7.

[18] Wang H, Suh JW, Das SR, Pluta JB, Craige C, Yushkevich PA. Multi-atlas segmentation with joint label fusion. IEEE Trans Pattern Anal Mach Intell 2013;35(3):611–23.

[19] Morra JH, Tu Z, Apostolova LG, Green AE, Avedissian C, Madsen SK, et al. Automated mapping of hippocampal atrophy in 1-year repeat {MRI} data from 490 subjects with Alzheimer's disease, mild cognitive impairment, and elderly controls. Neuroimage 2009;45(1 Suppl.):S3–15. Mathematics in Brain Imaging.

[20] Frankó E, Joly O. for the Alzheimer's Disease Neuroimaging Initiative. Evaluating Alzheimer's disease progression using rate of regional hippocampal atrophy. PLoS ONE 2013;8(8):e71354.