Physica Medica 30 (2014) 878–887

Contents lists available at ScienceDirect

Physica Medica

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

Original paper

Automated voxel-by-voxel tissue classification for hippocampal segmentation: Methods and validation

S. Tangaro [a], N. Amoroso [a, b, *], M. Boccardi [c], S. Bruno [d], A. Chincarini [e], G. Ferraro [a, b], G.B. Frisoni [c, f, g], R. Maglietta [h], A. Redolfi [c], L. Rei [e], A. Tateo [a], R. Bellotti [a, b], for the Alzheimers Disease Neuroimaging Initiative[1]

[a] Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Italy
[b] Dipartimento Interateneo di Fisica "M. Merlin", Università degli Studi di Bari, Italy
[c] LENITEM Laboratory of Epidemiology, Neuroimaging, and Telemedicine, IRCCS Centro S. Giovanni di Dio – FBF, Brescia, Italy
[d] Overdale Hospital, St Helier, Jersey
[e] Istituto Nazionale di Fisica Nucleare, Sezione di Genova, Italy
[f] AFaR Associazione Fatebenefratelli per la Ricerca, Rome, Italy
[g] Psychogeriatric Ward, IRCCS Centro S. Giovanni di Dio – FBF, Brescia, Italy
[h] Istituto di Studio sui Sistemi Intelligenti per l'Automazione, Consiglio Nazionale delle Ricerche, Bari, Italy

ARTICLE INFO

Article history:
Received 19 February 2014
Received in revised form
9 June 2014
Accepted 24 June 2014
Available online 11 July 2014

Keywords:
Hippocampus
Segmentation
MRI
Feature extraction
Classification methods

ABSTRACT

The hippocampus is an important structural biomarker for Alzheimer's disease (AD) and has a primary role in the pathogenesis of other neurological and psychiatric diseases. This study presents a fully automated pattern recognition system for an accurate and reproducible segmentation of the hippocampus in structural Magnetic Resonance Imaging (MRI). The method was validated on a mixed cohort of 56 T1-weighted structural brain images, and consists of three processing levels: (a) Linear registration: all brain images were registered to a standard template and an automated method was applied to capture the global shape of the hippocampus. (b) Feature extraction: all voxels included in the previously selected volume were characterized by 315 features computed from local information. (c) Voxel classification: a Random Forest algorithm was used to classify voxels as belonging or not belonging to the hippocampus. In order to improve the classification performance, an adaptive learning method based on the use of the Pearson's correlation coefficient was developed. The segmentation results (Dice similarity index = 0.81 ± 0.03) compare well with other state-of-the art approaches. A validation study was conducted on an independent dataset of 100 T1-weighted brain images, achieving significantly better results than those obtained with FreeSurfer.

Introduction

Magnetic resonance imaging (MRI) has acquired a primary role in both clinical practice and research, as it provides anatomical information on disease-related brain changes, useful for diagnostic purposes and treatment planning. The hippocampus is a brain structure with a key role in the pathophysiology of a number of common disorders, such as Alzheimer's disease (AD), schizophrenia and some forms of epilepsy. In AD in particular, the hippocampus loses volume at a faster rate than other brain regions, and it is therefore recognized as an important biomarker for the early diagnosis of the disease [1].

Until recently the segmentation of the hippocampus, *i.e.* its identification and separation from surrounding brain structures, had been performed mainly manually or with semi-automated techniques, followed by manual editing. This is obviously time-consuming and subject to investigator variability, so a number of automated segmentation methods have been developed. These have relied so far mainly on image intensity-based methods, often adopting multi-atlas registration approaches, in order to minimize errors due to individual anatomical variation. More recently,

* Corresponding author. Dipartimento Interateneo di Fisica "M. Merlin", Università degli Studi di Bari, Via Amendola 173, 70126 Bari, Italy.
E-mail address: nicola.amoroso@ba.infn.it (N. Amoroso).

[1] Data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the 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.2014.06.044
1120-1797/

though, a number of methods that exploit shape information have been developed, based on preliminary work carried out in the nineties with the Active Shape Models (ASM) and the Active Appearance Models (AAM) [2]. Recent work has employed probabilistic tree frames for brain segmentation, adopting in some cases specific models such as Markovian random fields or graphical cuts [3,4]. Alternative approaches involve, for example, machine learning techniques [5], labeling strategies combined with other methods such as multiple segmentations [6], longitudinal 4-D methods with graph cuts [7] and label fusion with template libraries [8]. In recent years, best accuracy levels have been achieved by template-warping methods incorporating label fusion strategies [9–11]. However, none of these methods has achieved widespread use.

This study classifies voxels in a brain MRI image as belonging or not belonging to the hippocampus, in order to create statistical map for the study of systematic differences between subjects with probable Alzheimers disease and healthy controls [12]. A Random Forests (RF) algorithm is used to classify voxels according to a number of features that describe complex images, based on a fully automated statistical analysis of adjacent groups of voxels [13]. A novel training procedure is proposed, involving the use of the Pearson's correlation coefficient between the test image and the training dataset, in order to select a "best fit" data subset for the training classification. This method has been called "active learning". In this way, two different procedures can be used, either a full training set learning phase (passive learning) or active learning. The latter results in better performance for the images with low correlation to the training dataset. The performances of the classifiers are measured by the following error metrics: Precision, Recall, Relative Overlap and Dice index. Also the use of active learning techniques reduces time-consuming computations of processing high-dimensional feature vectors thanks to the optimization of the training dataset.

Materials

The method was developed on a first dataset DB1, and subsequently tested on a second dataset DB2. DB1 consists of 56 T1-weighted whole brain images, and corresponding manually segmented bilateral hippocampi (masks), from the Laboratory of Epidemiology and Neuroimaging, IRCCS San Giovanni di Dio FBF in Brescia (Italy). The images included belonged to subjects with diagnoses of AD, mild cognitive impairment and normal controls. Twenty-nine of the subjects included in DB1 were affected by different cerebrovascular conditions or other disorders, referred to as "other".

All images were acquired on a Philips Gyroscan 1.0 T scanner according to the following parameters: gradient echo 3D technique, TR = 20 ms, TE = 5 ms, flip angle = 30°, field of view = 220 mm, acquisition matrix of 256 × 256, slice thickness of 1.3 mm, image dimensions 181 × 145 × 181. Detailed description of database is available in Ref. [14]. For the manual segmentation, the images were automatically re-sampled through an algorithm included in the MINC package (www.bic.mni.mcgill.ca/software) and normalized to the Colin27 template (www.bic.mni.mcgill.ca) with a voxel-size of 1.00 × 1.50 × 1.00 mm^3. When automated registration failed (about 5% of cases), manual registration was performed, based on 11 anatomical landmarks. Manual hippocampal segmentation were performed using the software Display 1.3 (www.bic.mni.mcgill.ca/ServicesSoftwareVisualization/Display), following the protocol defined in Ref. [15]. An external validation of the proposed novel algorithm was carried out on an independent dataset DB2 shared by the EADC-ADNI working group using a standard harmonized protocol (www.hippocampal-protocol.net). The more inclusive definition of the Harmonized protocol [16] may limit the inconsistencies due to the use of arbitrary lines and tissue exclusion of the currently available manual segmentation protocols. The mean Dice value for repeated manual segmentations on DB2 was 0.89 (range: 0.88–0.92). The Dice definition, as those of other error metrics adopted to assess the segmentation consensus, is provided in the Section 3.

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

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California – San Francisco. ADNI is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org.

Demographic data for DB1 and DB2 is shown in Table 1. The image processing and classification were carried out blindly to subject status.

Methods

In brief, the analysis consisted of the following main steps (see flow chart in Fig. 1):

- Volume of interest (VOI) extraction: this comprised a first step of atlas matching and registration, followed by further delimitation of the VOI through application of the new algorithm based on Point Distribution Model Theory;
- voxel-by-voxel feature extraction;
- voxel classification.

The performance of the segmentation method was subsequently assessed using the following metrics:

Table 1
Group size, range age (years) and sex of the two clinical datasets, containing normal control (NC) subjects, Alzheimer's disease (AD) and mild cognitive impairment (MCI) patients. Subjects affected by other cerebrovascular conditions are referred as "others".

Data	Size	Age	M/F	Subjects
DB-1	56	47–92	22/34	1 NC – 16 MCI – 10 AD – 29 other
DB-2	100	60–90	56/44	29 NC – 34 MCI – 37 AD

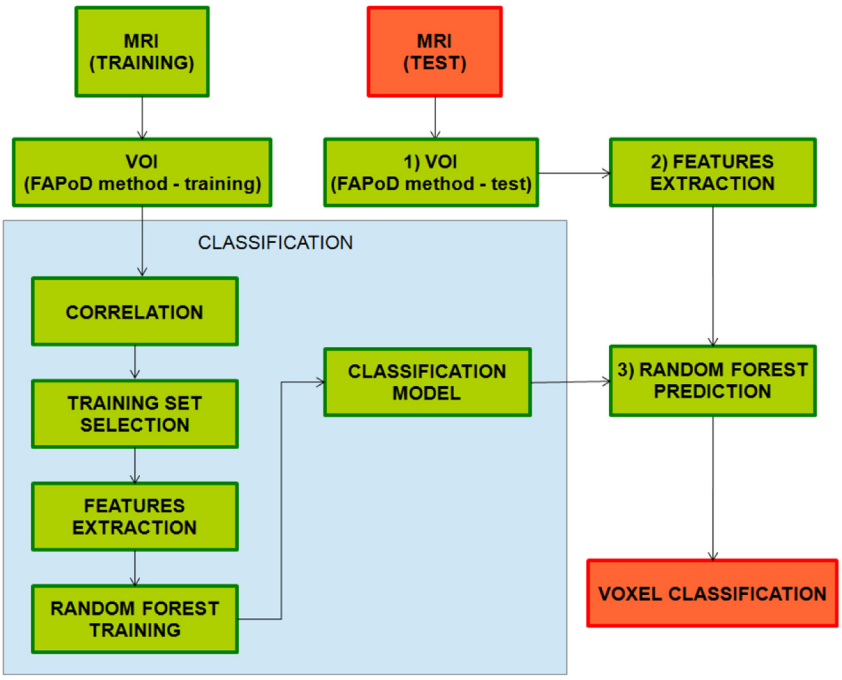

Figure 1. Flow chart of method, according to the following steps: 1) volume of interest (VOI) extraction, 2) determination of voxel features and 3) voxel classification trough Random Forests prediction model. Leave-one-out approach is used to test the performance of method. The training phase of active learning is shown in detail in the classification box. Red-boxes denote both the input and the output of test image. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

- Precision $= \frac{N(A \cap B)}{N(B)}$
- Recall $= \frac{N(A \cap B)}{N(A)}$
- Relative Overlap $= \frac{N(A \cap B)}{N(A \cup B)}$
- Dice $= \frac{2\,N(A \cap B)}{N(A)+N(B)}$

where A represents the set of manually segmented hippocampal voxels (ground truth segmentation), B the set of hippocampal voxels as segmented by the proposed algorithm (testing segmentation) [5] and N the relative number of elements.

Volume of interest extraction: FAPoD algorithm

First the images underwent preprocessing to be standardized in terms of intensity and spatial coordinates. A three dimensional noise-filtering pyramid is applied, followed by registration of the images to the Montreal Neurological Institute (*MNI*) standard template (ICBM152) using 12-parameter affine-registration, and subsequent re-sampling on an isotropic grid with 1 mm^3 voxel-size. After spatial normalization, intensity normalization was carried out too, producing images in which anatomical structures were substantially overlapping, and segmented in gray matter (GM), white matter (WM) and cerebro-spinal fluid (CSF), with remaining differences due to inter-individual variability and any disease-related effects [17]. This pre-processing was used only in registration phase.

Since all images were registered to the same template, the hippocampus was always localized approximately in the same region. The search volume could therefore be reduced by delimiting a VOI. This was done with a new method, called Fully Automatic Algorithm based on Point Distribution Model Theory (FAPoD) [18]. This method identifies a gross hippocampal boundary shape, improving the overall performance and reducing the computational cost of the analysis, because when testing a new image only the voxels inside the VOI are classified, and all others are assumed negative. The FAPoD model is a shape analysis algorithm that automatically and iteratively detects landmarks on a training dataset of manually segmented binary images to develop an accurate description of a mean hippocampal shape and its variability through a probabilistic map.

The landmark detection algorithm is shown in Fig. 2. For each image in the training set, the voxels belonging to the hippocampal edges (according to the manual segmentation) were identified, then the voxel occurrence frequency (*i.e.*, the number of times a voxel is associated with an edge in the training data) was calculated. For every coronal slice the highest frequency voxels were referred to as “reference voxels”. The reference voxels could not be used directly because they might not be present on each edge analyzed. For this reason, in every image edges, the nearest voxels to the reference voxels were identified as mathematical landmarks. Once mathematical landmarks had been fixed on every edge of binary image, pseudo-landmarks were defined as those whose distance from every chord subtended by two consecutive mathematical landmarks was the highest (Fig. 3).

First, the number of reference voxels necessary for the boundary reconstruction for each coronal projection, was identified. This depended on the mean length of the contour of the hippocampus. The minimum number of landmarks necessary for the reconstruction of an edge was 3. Several tests showed that a number of landmark greater than 8 worsened the reconstruction for the edges of larger dimensions, according to this criterion 4 mathematical landmarks and 4 pseudo-landmarks were adopted. The interpolation according to the landmark variability yields a super-imposition of different possible contours. The number of super-imposed contours can be interpreted as the probability of a considered voxel to be a contour voxel, and in this way a probability map is constructed. Inner hippocampal voxels have higher probability, belonging to a greater number of possible contours, whilst the probability decreases towards the background voxels. According to the Point

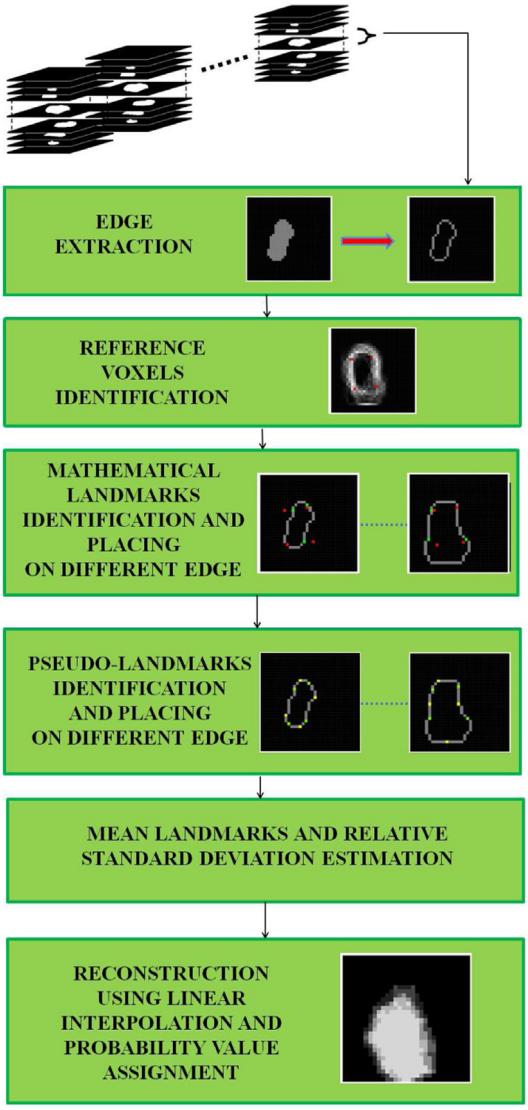

Figure 2. Description of the volume of interest extraction method. The mathematical landmarks are individuated in the following way: the voxels belonging to training hippocampi (according to manual segmentation) with higher frequency are computed (reference voxels in red); next the mathematical (green) and the pseudo-landmarks (yellow) are recognized onto every image contour.

Distribution Model Theory the mean hippocampal shape can be described by the mean and standard deviation for each landmark series. These mean landmarks and the respective standard deviations landmarks σ_l were combined with a pointwise linear interpolation generating a probability map.

The, the number of images useful to retrieve the FAPoD volume was established. An evaluation index is the volume of interest measure. Figure 4 shows how the volume obtained by FAPoD methods plotted versus the number of images used plateaus using about 25 images considering $3\sigma_l$, $2\sigma_l$ or $1\sigma_l$ for both left and right hippocampi.

Feature extraction

Supervised pattern recognition systems involve taking a set of labeled examples (features) and learning a pattern based on those examples. The features should contain information relevant to the

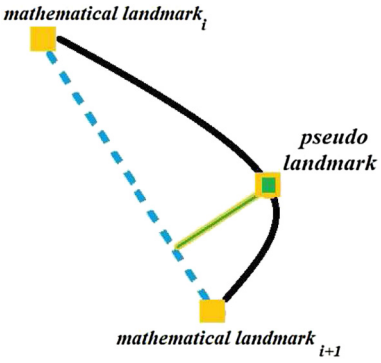

Figure 3. Once mathematical landmarks have been fixed on every labeled hippocampus, pseudo-landmarks are defined as those whose distance from every chord subtended by two consecutive mathematical landmarks is the highest.

classification task. In the analysis presented here, for each voxel, a vector whose elements represent information about position, intensity, neighboring texture, and local filters was obtained. Texture information (contrast, uniformity, rugosity, regularity, etc.) was expressed using Haar-like and Haralick features, as in recent work on automated segmentation [5]. These features are characterized by computational simplicity: for each voxel, a value obtained by the weighted sum of the intensities on the area spanned by a template, the sum of the weights being zero, was extracted. Filters of size varying from $3 \times 3 \times 3$ to $9 \times 9 \times 9$ were used for the calculation of Haar-like features.

The Haralick features were calculated from the normalized gray level co-occurrence matrices (GLCM) created on the m $\times$ m voxel projection subimages of the volume of interest. For each voxel values of m varying from 3 to 9 were used. Within each co-occurrence matrix M an element p_{ij} represents an estimate of the probability that two voxels with a specified polar separation (d, θ) have gray levels i and j. Coordinates d and θ are, respectively, the distance and the angle between the two voxels i and j. In the present work $d = 1$ and the displacements at quantized angles $\theta = k\pi/4$, with $k = 0,1,2,3$ were considered. As shown in other papers a subset of Haralick feature is sufficient to obtain satisfactory performance for discrimination problem, as in Refs. [19,20]. Preliminary recognition experiments individuated 4 Haralick features giving the best recognition rate:

- energy:

$$f_1 = \Sigma_{ij} p_{ij}^2 \tag{1}$$

- contrast:

$$f_2 = \Sigma_{n=0}^{N_g-1} n^2 \Sigma_i^{N_g} \Sigma_j^{N_g} p_{ij}^2; \left| i - j \right| = n \tag{2}$$

- correlation:

$$f_3 = \frac{\Sigma_{ij}(ij) p_{ij}^2 - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{3}$$

where μ_x, μ_y, σ_x, and σ_y σ_y are the means and standard deviations of p_x and p_y, the partial probability density functions obtained summing the rows or the columns of p_{ij}.

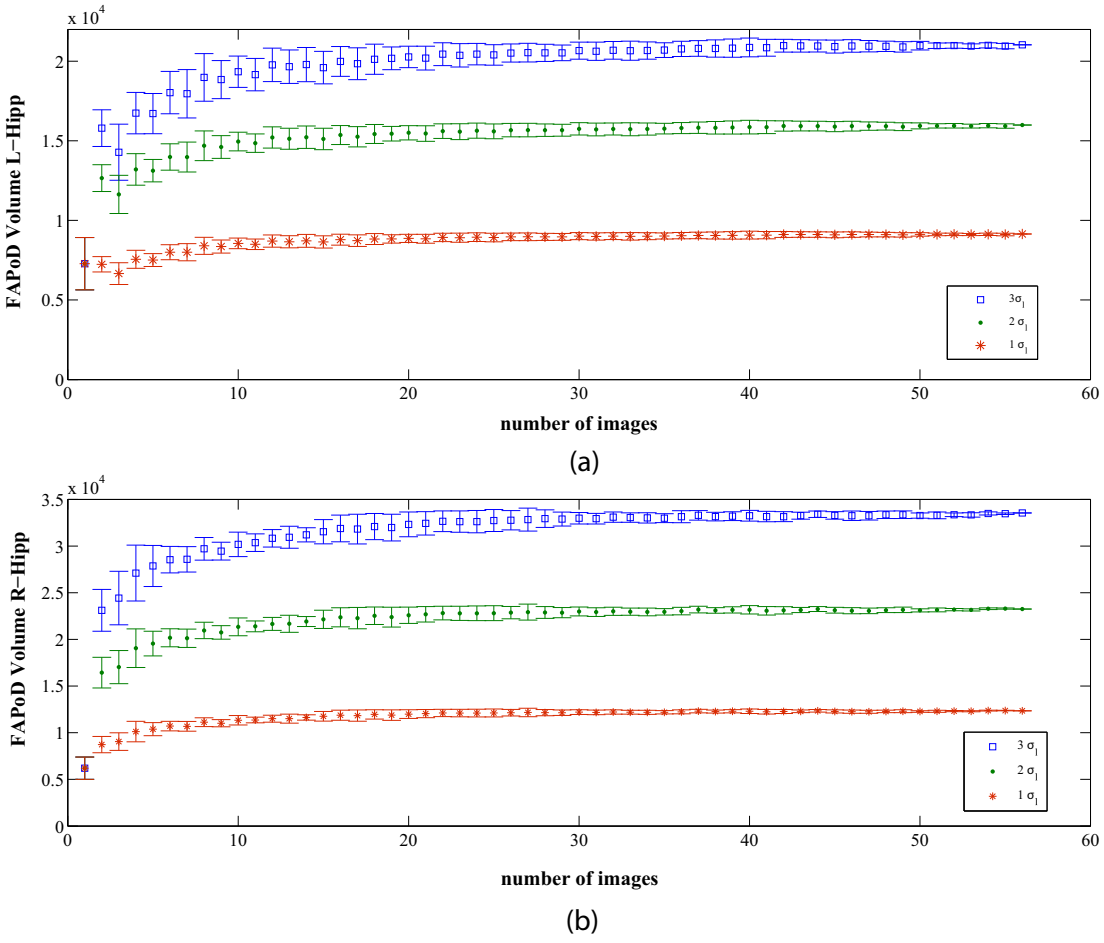

Figure 4. The volume reconstructed by FAPoD (in mm^3) varying the number of images used to retrieve the volume for (a) left and (b) right hippocampus.

- inverse difference moment:

$$f_4 = \Sigma_{ij} \frac{p_{ij}}{1 + (i-j)^2} \tag{4}$$

Finally, for each voxel, the gradients were calculated in all directions at distances of one voxel, and the relative positions of the voxels, *x*, *y* and *z*, were included as additional features. The best analysis configuration, expressed by highest metrics mean value, was obtained with 315 features.

Voxel classification

Due to the large number of discriminant features used to describe the VOI, a number of different multivariate classification algorithms were considered [21,22]. RF was in the end selected, for its better performance, even in the presence of data with high variability and noise. RF consists of a collection of tree classifiers, built on a sample of training data obtained by bootstrap technique. The final output of the classifier is calculated through the majority of the votes among the individual classifiers. In this work the forests were used to classify voxels as belonging or not to the hippocampus. A vector containing all features values plus the target value was assigned to each voxel (*1* if the voxel belonged to the hippocampus and *0* if it did not). With regard to the number of trees *N*, RF of different sizes, even of 1000 trees, were explored, but the out-of-bag error was rather constant from 100 trees onward. The agreement between the actual labels and the predictions was calculated using the error metrics described in the method section.

Active learning is a new method for training dataset selection during the learning phase of classification, based on the assumption that some examples may hinder the learning process rather than aid it. By focusing on informative examples, the rate of learning is optimized, aiding the search through the mathematical space of classification hypotheses. The Pearson's coefficient *r* on gray levels between the target image and each image in the training dataset was used as a measure of the degree in which each example is informative in relation to the target image. For each target image the distribution of Pearson's coefficient was obtained and the mean value $\bar{r}$ was calculated. Figure 5 shows the distribution of $\bar{r}$ values. Each entry is the average value of the Pearson's coefficient distribution on gray levels between the target image and each image in training dataset.

For each image the classification behavior has been studied depending on the distribution of $\bar{r}$ as illustrated in Fig. 5. For those images (in validation) with mean value $\bar{r} \geq \bar{r}_{th}$ (where $\bar{r}_{th}$ is a threshold value corresponding to the first quartile of the $\bar{r}$ distribution) the whole remaining training dataset was used (passive learning). For the other cases only the most correlated images were selected (active learning).

In the latter instance using the whole database would not be as efficient, because the informative power of the labeled images is

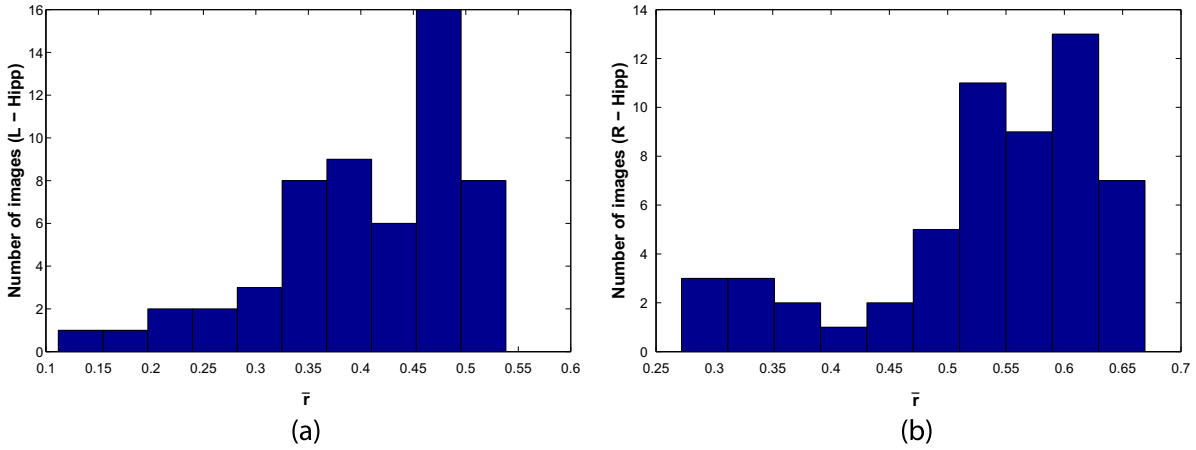

Figure 5. Distribution of $\bar{r}$ values for left (a) and right (b) hippocampus. Each entry is the average value of the Pearson's coefficient distribution on gray levels between the target image and each image in training dataset.

diminished by the presence of misleading examples. With active learning the dataset is reduced, but more adherent to the image to be segmented, which improves the classification performance. In fact, for each particular image to be segmented, the correlation coefficient between such image and the training images is proportional to the probability of that image to be included in the training dataset. The Pearson's coefficient measures the strength and direction of linear relationship. In this way the training dataset includes only the images most correlated with the image to be tested, resulting in the construction of a specific classifier for each hippocampus to be segmented. To determine the optimal number of correlated images for active learning, the segmentation performance on DB1 was studied as a function of the correlation. The most correlated images were defined as those exceeding the 95%, 90%, 85%,…,50% of the maximum correlation value. We found that performances reached a plateau around a correlation value equal to the 90%. Therefore only those images were included in the reduced dataset. The system is shown schematically in the classification box of Fig. 1.

Computational infrastructure

The analyses presented in this paper were developed in MATLAB framework. The algorithms required substantial computational resources, with segmentation times of about 1 h per image. Therefore the availability of distributed computing software environments and adequate infrastructures was of fundamental importance.

In this study, the LONI pipeline processing environment was used, a user-friendly and efficient software for complex data analyses, available at http://pipeline.loni.usc.edu.

The study was carried out using the local computer farm BC2S (http://www.recas-pon.ba.infn.it), a distributed computing infrastructure consisting of about 5000 CPU and allowing up to 1,8 PB storage. A further study for grid deployment was also performed, with the aim of creating a pipeline tool suitable for large clinical trials. It was carried out on the European Grid Infrastructure (EGI), which consists of about 300 geographically distributed sites around the world. In particular all the results presented in this study were obtained on the BC2S using the 56 MR images at our disposal. In this context a feasibility study concerning preprocessing analyses (brain extraction and registration) on about 3000 simulated MRI images was performed. The results in terms of run-time and failure rate for every submitted job were reported in Ref. [23].

Results

Results on DB1

When validating a machine learning approach it is essential to examine error metrics on both the training and testing sets. A test set independent of the training set is of fundamental importance to show the effectiveness of a classifier. Since 56 hand-labeled brains were used to train the algorithm, a leave-one-out analysis was employed to ensure a separation between the training and testing sets.

The optimal number n of standard deviation σ_l to be considered for FAPoD volume retrieval was carried out by evaluating the performance of the overall system. Using a leave-one-out configuration on the whole training dataset (passive learning) performances were comparable for $n = 3$, 2 and 1 as shown in Fig. 6; the $2\sigma_l$ confidence interval was preferable because it represented a good compromise between the number of false positives, false negatives and the computational efficiency of the algorithm (for each hippocampus the average of the voxel labeled not considered is less than 1%). The relative Dice coefficient was 0.81 ± 0.04 (mean value ± standard deviation) for left hippocampi and 0.80 ± 0.05 for right hippocampi.

The correlation between each target image and training dataset was then investigated and it was verified that all images with lower values in $\bar{r}$ distribution had poor performances. Therefore, for these target images, a reduced training set was considered to improve the performances, while it was not useful in cases with higher $\bar{r}$ value. Figure 7 illustrates the behavior of the Dice index for images included in each quartile of $\bar{r}$ distribution. The four points are the mean Dice for images belonging to the four quartiles (0.00–0.25, 0.025–0.05, 0.05–0.75 and 0.75–1.00) sampled from the $\bar{r}$ distribution shown in Fig. 5. Group-wise average differences were significant (t-test < 0.05). The reduced training dataset was useful only for images with $\bar{r}$ value under the lower quartile of $\bar{r}$ distribution; in the other cases the whole training dataset was used. The maximum improvement using a reduced training dataset was obtained for the image with the greatest difference between the mean of Pearson's coefficient of the reduced and the whole training dataset: 7% for Precision; 6% for Recall; 9% for Relative Overlap; 7% for Dice index.

Using a reduced training dataset only for the images whose $\bar{r}$ value was in the lower quartile of $\bar{r}$ distribution we found an average Dice index of 0.81 ± 0.03 (for left hippocampus) and of 0.80 ± 0.05 (for right hippocampus). The active learning method

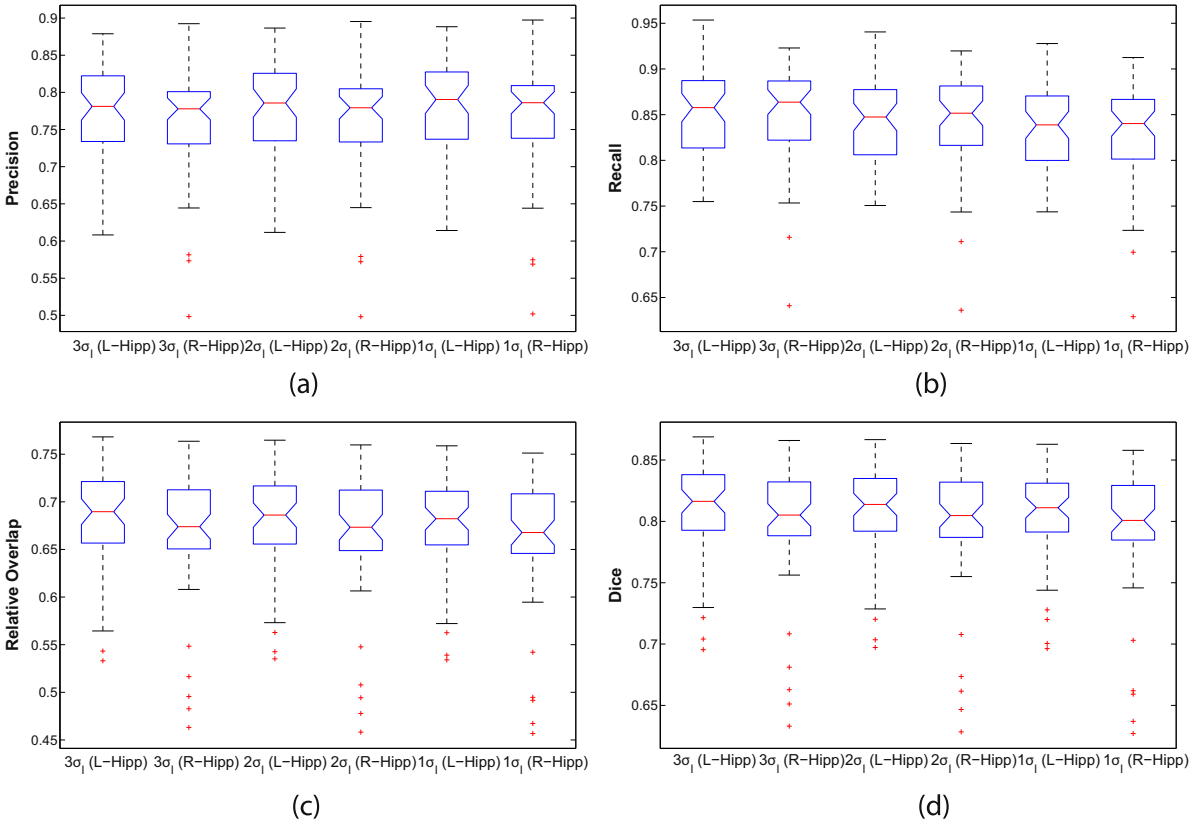

Figure 6. Classification comparison: (a) Precision, (b) Recall, (c) Relative Overlap and (d) Dice index show the presented algorithm performance varying the number of σ_I of FAPoD method (passive learning) for left and right hippocampus (L-Hipp, R-Hipp). Boxes have lines at the lower quartile, median, and upper quartile values, with whiskers extending to 1.5 times the inter-quartile range. Outliers are indicated by a plus sign.

compared well with the passive learning approach offering an improvement for those images poorly correlated with the training set without a significant worsening of the computational burden. As a consequence the active learning approach was used for a further test on an independent dataset.

Segmentation differences were also qualitatively addressed. Figure 8 shows examples of super-imposition between the manual segmentation (in blue) and automated segmentation (in red) from the test set. There is good differentiation of the hippocampus from the surrounding structures (amygdala, CSF, and adjacent white matter), and the tracings are smooth, and similar to those obtained with manual segmentation. The segmentation errors appear to be uniformly distributed among the head, the body and the hippocampal tail. This image is representative of the segmentation accuracy obtainable on the test images.

Results on DB2

In the last stage of the study, the new method described here was compared with the publicly available brain segmentation package FreeSurfer v.5.1 (FS) [24]. At this aim a second experiment was conducted on an independent dataset. The results are shown in Table 2 and Fig. 9. The presented algorithm compares well with FS in terms of the previously defined error metrics.

The hippocampal volumes obtained by the proposed algorithm confirm the hippocampal atrophy as a supportive feature for the AD diagnosis. For left (right) hippocampal volumes (in mm^2) we found $V_{NC} = 4040\pm622$, $V_{MCI} = 3345\pm738$ and $V_{AD} = 3134\pm654$ ($V_{NC} = 3874\pm869$, $V_{MCI} = 3212\pm728$ and $V_{AD} = 3035\pm1068$) to be compared with the manually labeled left (right) volumes $V_{NC} = 4125\pm483$, $V_{MCI} = 3558\pm493$ and $V_{AD} = 3305\pm598$ ($V_{NC} = 4091\pm467$, $V_{MCI} = 3591\pm527$ and $V_{AD} = 3429\pm653$). The NC, MCI and AD populations were found significantly different with a Kruskal–Wallis test ($p < 0.01$) performed on the segmentation volumes.

Discussion and conclusion

In recent years, the development of neuroimaging and signal processing has made possible the visualization and measurement of pathological brain changes in vivo, producing a radical change, not only in the field of scientific research, but also in the everyday clinical practice [25]. The development of tools for reliable and accurate anatomical segmentation is of crucial importance for the quantitative analysis of brain images. This work proposes an automated method for the segmentation of images of the hippocampus in brain MRI.

This is an innovative approach based on the use of discriminating features and on their classification by means of a RF classifier in a VOI defined using the new FAPoD method. This method, only based on shape evaluations, was able to deal with the database heterogeneity within a $2\sigma_I$ variation. Future improvement of the method could be obtained combining shape-based and intensity based information, perhaps using warping methods before the application of FAPoD.

To the best of our knowledge, this is the first application of a RF classifier to hippocampal segmentation combined with expert priors on shape. The number of features required for a robust classification is here much lower than in other studies Ref. [5], which however show slightly better performances (Dice index = 0.85).

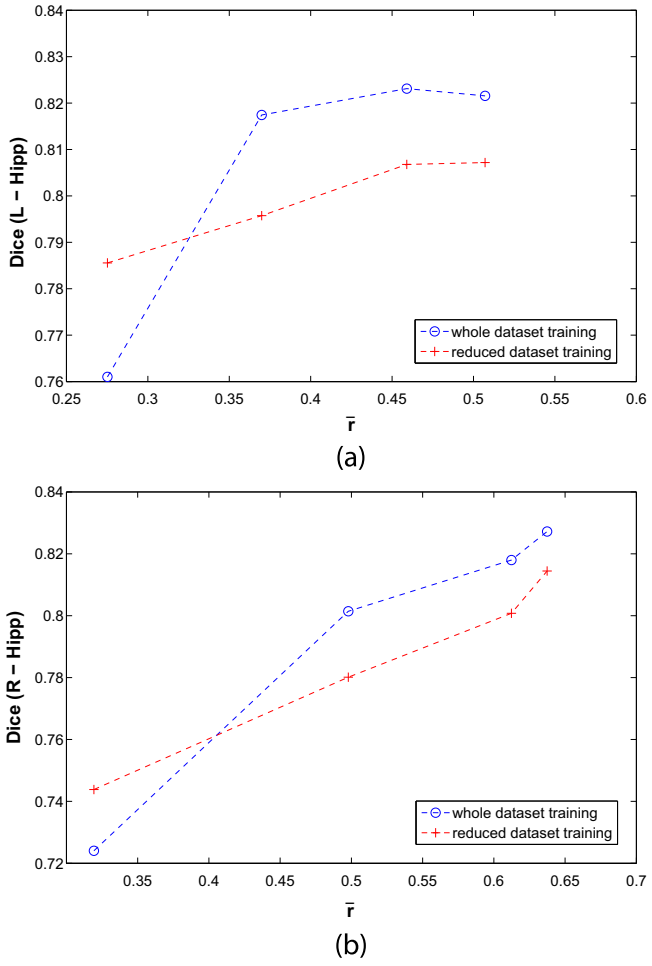

Figure 7. Behavior of the Dice index for images included in each quartile of $\bar{r}$ distribution is shown for both left (a) and right (b) hippocampi. Average values for $\bar{r}$ in each quartile of $\bar{r}$ distribution versus average Dice index values are plotted. The group mean differences were obtained with 0.05 significance level.

Previous studies, using atlas-based approaches, reported Dice coefficients in the range 0.75–0.88 for both healthy and mixed cohorts (*i. e.* cohorts composed of both healthy controls and diseased subjects). The best results in the literature have been so far acquired through patch-based multi-atlas segmentations [26,9–11]. The average Dice index of 0.81 ± 0.03 obtained in this study is comparable to existing results for mixed cohorts. Besides it is worthwhile to note that the performance of the presented method was tested on a 1.0 T dataset, which suffers from a lower signal-to-noise ratio compared with high-field datasets. Another important feature to be outlined when comparing segmentation performances concerns the dependence of the segmentation results with the segmentation protocol used for the manual labeling [27]. The active learning in the classification step improves the performance of the method particularly for images with $\bar{r} < \bar{r}_{th}$, *i.e.* active learning is useful for data poorly correlated with the training set (about 15% of images of our database). The improvement using a reduced training dataset was about 0.01–0.07 for Dice index, depending on the difference between the mean of Pearson's coefficient of the reduced and the whole training dataset. This improvement in our database covers approximately 10% of the images.

A second experiment conducted on an independent dataset of 100 T1-weighted structural brain images showed that this method is able to perform significantly better than FreeSurfer, in terms of reliability (similarity index = 0.76 ± 0.07 vs 0.71 ± 0.09), these results are confirmed for the other error metrics (precision = 0.84 ± 0.06, recall = 0.71 ± 0.09, relative overlap = 0.62 ± 0.08) resulting in higher values than FreeSurfer (precision = 0.73 ± 0.09, 0.68 ± 0.10, 0.56 ± 0.08).

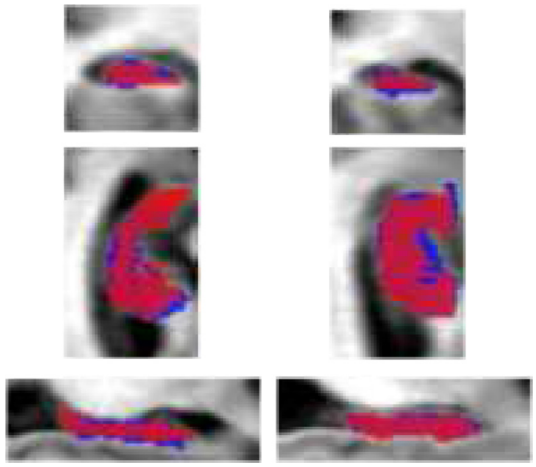

Figure 8. Examples of automated segmentation (in red) for one of the worst cases (left) and one of the best cases (right) from the testing set. The manual labeling is in blue. The coronal, axial, and sagittal views are shown respectively from top to bottom. Mislabels tend to be uniformly distributed. For the example of poor automated segmentation major issues arise in the neighboring region between hippocampus and amygdala.

Differences in image quality, manual segmentation protocol, clinical status and demographics have been described as possible causes of discrepancy between different clinical datasets. The poor relative overlap, which may indicate a substantial difficulty for the algorithms to reproduce manual segmentation, is probably related the intrinsic variability involved in manual segmentation. The currently available protocols for manual segmentation include features and information that are not entered in the training of automated algorithms, generating additional source of variability. The inclusion or exclusion of hippocampal white matter, the use of arbitrary lines, the exclusion of parts of the hippocampal tail or of hippocampal gray matter held as vestigial [28] lead to non-systematic differences in hippocampal segmentations across subjects. In fact, a critical point of the use of machine learning algorithms is the possibility of using a training dataset, with a large number of examples, shared by the scientific community, in line with the efforts towards a standard harmonized protocol by the EADC-ADNI working group [29]. In the case of medical images the creation of large databases is very time-consuming and methodologically challenging. The results here presented suggest the need for further efforts to achieve a universally recognized gold standard, both to accurately compare hippocampal

Table 2
Mean and standard deviation metric values on test dataset (independent from training dataset) are shown. The group mean differences were obtained with 0.01 significance level. Moreover, the segmented volumes for both left and right hippocampi are shown.

	Presented method		FreeSurfer	
	Left hipp	Right hipp	Left hipp	Right hipp
Precision	0.84 ± 0.06	0.84 ± 0.06	0.73 ± 0.09	0.75 ± 0.09
Recall	0.71 ± 0.09	0.72 ± 0.07	0.68 ± 0.10	0.68 ± 0.10
Relative overlap	0.62 ± 0.08	0.64 ± 0.06	0.55 ± 0.08	0.56 ± 0.08
Dice	0.76 ± 0.07	0.74 ± 0.05	0.71 ± 0.09	0.71 ± 0.09
Volumes (mm^3)	3350 ± 743	3526 ± 944	3502 ± 875	3678 ± 889

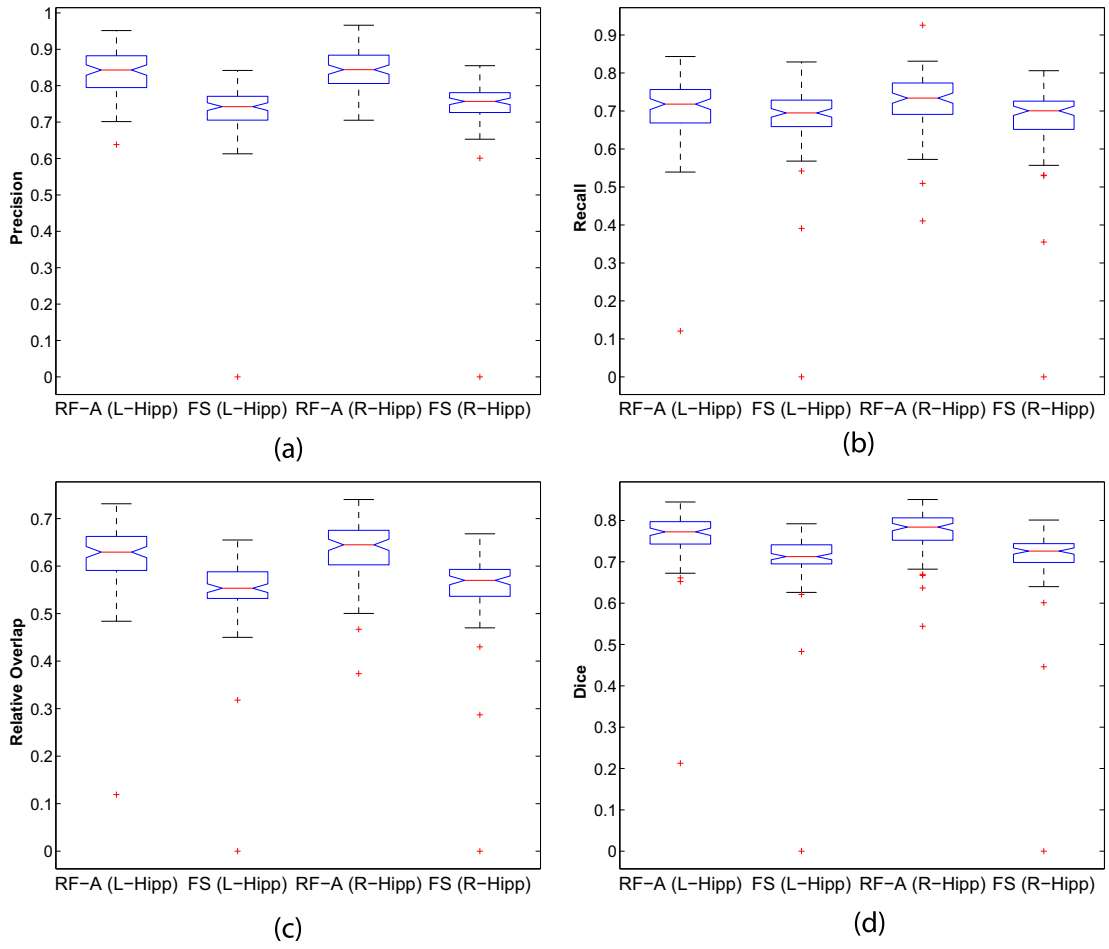

Figure 9. The figure shows the comparison between the FreeSurfer and the proposed algorithm performances on the test sets: (a) Precision, (b) Recall, (c) Relative Overlap and (d) Dice index show that the classification performances of the described method on the test set is definitively better than those obtained in training.

segmentation algorithms and, above all, to improve the robustness clinical classification.

This approach could be further explored in the future using different evaluation coefficients to measure the similarity or the distance between the image to be segmented and the training dataset. Also non-linear relationships to select training dataset will be investigated. Active learning is a “data-oriented” approach, which could play a fundamental role in the use of distributed computing infrastructures by reducing the training set size and, therefore, overcoming upload/download problems and reducing the training phase time.

Other possible future developments of the method include the application of active learning to voxel features and the use of non-linear registration algorithms, to further improve segmentation. The performance of this approach could be further investigated on larger clinical databases using publicly available data and advanced distributed virtual laboratories [30].

Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (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 the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

This research was supported by Istituto Nazionale di Fisica Nucleare (INFN), Italy. This research was also supported 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.

References

[1] Frisoni GB. Structural imaging in the clinical diagnosis of alzheimer's disease: problems and tools. J Neurol Neurosurg Psychiatry 2001;70(6):711–8.
[2] Cootes TF, Edwards GJ, Taylor CJ. Active appearance models. IEEE Trans Pattern Anal Mach Intell 2001;23(6):681–5.
[3] Zhang Y, Brady M, Smith S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans Med Imag 2001;20(1):45–57.
[4] Song Z, Tustison N, Avants B, Gee JC. Integrated graph cuts for brain MRI segmentation. Proc Med Image Comput Comput Assist Interv 2006;9(2):831–8.
[5] Morra JH, Tu Z, Apostolova LG, Green AE, Avedissian C, Madsen SK, et al. The Alzheimer's disease neuroimaging initiative, validation of a fully automated 3D hippocampal segmentation method using subjects with Alzheimer's disease mild cognitive impairment, and elderly controls. NeuroImage 2008;43(1):59–68.
[6] Heckemann RA, Hajnal JV, Aljabar P, Rueckert D, Hammers A. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage 2006;33(1):115–26.
[7] Wolz R, Aljabar P, Hajnal JV, Hammers A, Rueckert D. The Alzheimer's disease neuroimaging initiative, LEAP: learning embeddings for atlas propagation. NeuroImage 2010;49(2):1316–25.
[8] Collins DL, Pruessner JC. Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting animal with a template library and label fusion. NeuroImage 2010;52(4):1355–66.
[9] Coupé P, Manjón JV, Fonov V, Pruessner J, Robles M, Collins DL. Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. NeuroImage 2011;54(2):940–54.
[10] Kim H, Mansi T, Bernasconi N, Bernasconi A. Surface-based multi-template automated hippocampal segmentation: application to temporal lobe epilepsy. Med Image Anal 2012;16(7):1445–55.
[11] K. Kwak, U. Yoon, D. Lee, G. H. Kim, S. W. Seo, D. L. Na, et al. Fully-automated approach to hippocampus segmentation using a graph-cuts algorithm combined with atlas-based segmentation and morphological opening. Magnetic resonance imaging.
[12] Chincarini A, Bosco P, Gemme G, Morbelli S, Arnaldi D, Sensi F, et al. Alzheimer's disease markers from structural MRI and FDG-PET brain images. Euro Phys J Plus 2012;127(11):135.
[13] Cascio D, Magro R, Fauci F, Iacomi M, Raso G. Automatic detection of lung nodules in ct datasets based on stable 3d mass–spring models. Comput Biol Med 2012;42(11):1098–109.
[14] Sabattoli F, Boccardi M, Galluzzi S, Treves A, Thompson PM, Frisoni GB. Hippocampal shape differences in dementia with lewy bodies. Neuroimage 2008;41(3):699–705.
[15] Pruessner JC, Li LM, Serles W, Pruessner M, Collins DL, Kabani N, et al. Volumetry of hippocampus and amygdala with high-resolution MRI and three-dimensional analysis software: minimizing the discrepancies between laboratories. Cereb Cortex 2000;10:433–42.
[16] Boccardi M, Bocchetta M, Apostolova L, Barnes J, Bartzokis G, Corbetta G, et al. DelphiDelphi consensus on landmarks for the manual segmentation of the Hippocampus on MRI: preliminary results from the EADC-ADNI harmonized protocol working group. Neurology 2012;78(Suppl. 1):171–4.
[17] Chincarini A, Bosco P, Calvini P, Gemme G, Esposito M, Olivieri C, et al. The Alzheimer's disease neuroimaging initiative, local MRI analysis approach in the diagnosis of early and prodromal Alzheimer's disease. NeuroImage 2011;58(2):469–80.
[18] Amoroso N, Bellotti R, Bruno S, Chincarini A, Logroscino G, Tangaro S, et al. Automated shape analysis landmarks detection for medical image processing. Proc Int Symp CompIMAGE 2012:139–42.
[19] Bellotti R, De Carlo F, Gargano G, Maggipinto G, Tangaro S, Castellano M, et al. A completely automated CAD system for mass detection in a large mammographic database. Med Phys 2006;33(8):3066–75.
[20] Tangaro S, De Carlo F, Gargano G, Bellotti R, Bottigli U, Masala GL, et al. Mass lesion detection in mammographic images using Haralik textural features. Proc Int Symp CompIMAGE 2006-Comput Model Objects Represent Imag Fundam Meth Appl 2007:429–34.
[21] Maglietta R, Amoroso N, Bruno S, Chincarini A, Frisoni GB, Inglese P, et al. Random forest classification for hippocampal segmentation in 3D MR images, IMLA; 2013. conference inproceedings.
[22] S. Tangaro, N. Amoroso, S. Bruno, A. Chincarini, G. B. Frisoni, R. Maglietta, et al. Active learning machines for automatic segmentation of Hippocampus in MRI. Industrial Conference in Data Mining (ICDM 2013). LECTURE NOTES IN COMPUTER SCIENCE.
[23] N. Amoroso, M. Antonacci, M. Boccardi, M. Bocchetta, A. Chincarini, D. Diacono, et al. MRI analysis on a grid-based infrastructure using LONI pipeline, submitted on methods of Information in Medicine.
[24] Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neurotechnique 2002;33:341–55.
[25] Bellotti R, Pascazio S. Editorial: advanced physical methods in brain research. Eur Phys J Plus 2012;127(11):145.
[26] M. J. Cardoso, K. Leung, M. Modat, S. Keihaninejad, D. Cash, J. Barnes, et al. Steps: similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation. Medical image analysis.
[27] Nestor SM, Gibson E, Gao F, Kiss A, Black SE. A direct morphometric comparison of five labeling protocols for multi-atlas driven automatic segmentation of the hippocampus in alzheimer's disease. Neuroimage 2012;66(1):50–70.
[28] Boccardi M, Ganzola R, Bocchetta M, Pievani M, Redolfi A, Bartzokis G, et al. Defining survey of protocols for the manual segmentation of the Hippocampus: preparatory steps towards a joint EADC-ADNI harmonized protocol. J Alzheimer's Dis 2011;26(Suppl. 3):61–75.
[29] Frisoni GB, Jack CR. Harmonization of magnetic resonance-based manual hippocampal segmentation: a mandatory step for wide clinical use. Alzheimers Dement 2011;7(2):171–4.
[30] Frisoni GB, Redolfi A, Manset D, Rousseau M, Toga A, Evans AC. Virtual imaging laboratories for marker discovery in neurodegenerative diseases. Nat Rev Neurol 2011;7:429–38.