Hippocampal unified multi-atlas network (HUMAN): protocol and scale validation of a novel segmentation tool

2015 Phys. Med. Biol. 60 8851

(http://iopscience.iop.org/0031-9155/60/22/8851)

OPEN ACCESS
IOP Publishing | Institute of Physics and Engineering in Medicine Physics in Medicine & Biology
Phys. Med. Biol. **60** (2015) 8851–8867 doi:10.1088/0031-9155/60/22/8851

Hippocampal unified multi-atlas network (HUMAN): protocol and scale validation of a novel segmentation tool

N Amoroso[1,2], R Errico[1,3], S Bruno[4], A Chincarini[3], E Garuccio[5], F Sensi[3], S Tangaro[2], A Tateo[2], R Bellotti[1,2] and for the Alzheimers Disease Neuroimaging Initiative[6]

[1] Dipartimento Interateneo di Fisica 'M Merlin', Università degli Studi di Bari 'Aldo Moro', 70121 Bari, Italy
[2] Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Italy
[3] Istituto Nazionale di Fisica Nucleare, Sezione di Genova, Italy
[4] Overdale Hospital, St Helier, Jersey JE1 3UH, UK
[5] Dipartimento di Fisica, Università degli Studi di Siena, 55, 53100 Siena SI, Italy

E-mail: sonia.tangaro@ba.infn.it

Received 7 April 2015, revised 3 September 2015
Accepted for publication 28 September 2015
Published 4 November 2015

Abstract

In this study we present a novel fully automated Hippocampal Unified Multi-Atlas-Networks (HUMAN) algorithm for the segmentation of the hippocampus in structural magnetic resonance imaging. In multi-atlas approaches atlas selection is of crucial importance for the accuracy of the segmentation. Here we present an optimized method based on the definition of a small peri-hippocampal region to target the atlas learning with linear and non-linear embedded manifolds. All atlases were co-registered to a data driven template resulting in a computationally efficient method that requires only one test registration. The optimal atlases identified were used to train dedicated artificial neural networks whose labels were then propagated and fused to obtain the final segmentation. To quantify data heterogeneity and protocol inherent effects, HUMAN was tested on two independent data sets provided by the Alzheimer's Disease Neuroimaging Initiative and the Open Access

[6] 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

Series of Imaging Studies. HUMAN is accurate and achieves state-of-the-art performance ($\text{Dice}_{\text{ADNI}} = 0.929 \pm 0.003$ and $\text{Dice}_{\text{OASIS}} = 0.869 \pm 0.002$). It is also a robust method that remains stable when applied to the whole hippocampus or to sub-regions (patches). HUMAN also compares favorably with a basic multi-atlas approach and a benchmark segmentation tool such as FreeSurfer.

Keywords: hippocampus segmentation, machine learning, multi-atlas

Online supplementary data available from stacks.iop.org/PMB/60/8851/mmedia

(Some figures may appear in colour only in the online journal)

1. Introduction

The hippocampus is a brain structure of great importance for the pathogenesis of a number of neurodegenerative diseases. Hippocampal atrophy is an established primary biomarker in Alzheimer's disease (Sabuncu *et al* 2011, Chincarini *et al* 2013). The gold standard for hippocampal segmentation is manual tracing, which is time consuming, and subject to protocol and rater variability. This, along with the intrinsic difficulty of the task, has generated the need for automated segmentation techniques. Several methodologies have been put forward, including the state-of-the-art multi-atlas approaches, which are based on the non-linear co-registration of the target image with expert-segmented examples (*atlases*).

Several studies have demonstrated that multi-atlas accuracy is significantly related to the 'similarity' between the target image and the training atlases (Aljabar *et al* 2009, Lötjönen *et al* 2010, Kim *et al* 2013, Kwak *et al* 2013), but an objective definition of the optimal similarity is lacking. In the initial studies such similarity was based on demographic and intensity based criteria after linear (Leung *et al* 2010) or a non-linear registration (Klein *et al* 2008). More recently, non parametric manifold strategies, such as Isomap or Laplacian Eigenmaps, were investigated for atlas selection (Wolz *et al* 2010, Duc *et al* 2013). However, in some cases, parametric techniques, such as the Stochastic Neighbor Embedding, perform better than non-parametric ones (Van der Maaten and Hinton 2008). Overall, it is fair to say that an optimal atlas selection strategy is yet to be established, which is why we performed a comparison of different strategies to evaluate their effectiveness.

Multi-atlas approaches have some intrinsic drawbacks. First of all, errors during the registration phase, in the warp estimation or the label resampling can limit the reliability of the results (Pipitone *et al* 2014). In general, registration strategies incorporating tissue classification information can limit these issues, at the expense of increased processing times (Heckemann *et al* 2010). In addition, as multi-atlas accuracy depends on the similarity between training and test sets, large training sets ('complete' in a mathematical sense), requiring a vast amount of computational resources, are necessary to avoid poor performance. In principle, machine learning approaches can overcome these issues by generalizing the models learned by training samples. However, so far classification-based approaches (Morra *et al* 2010, Maglietta *et al* 2015, Tangaro *et al* 2013, 2014) have not attained performances comparable to multi-atlas methods. An effective combined strategy would seem a natural and elegant solution, as suggested by recent work (Wang *et al* 2011, Hao *et al* 2014). Interestingly, these studies show how voxel-wise learning can effectively introduce shape or context information in the segmentation process, improving its accuracy. Nonetheless, they have focused on label fusion, a particular aspect of multi-atlas approaches.

In general the comparison of different segmentation algorithms is arduous, due to the fact that most studies have different data sources or validation techniques. Also, when dealing

specifically with hippocampal segmentation the differences in segmentation protocols represent a particularly limiting constraint (Bellotti and Pascazio 2012, Bruno *et al* 2012, Nestor *et al* 2013). Although in recent years a considerable effort has been invested in the creation of a unified segmentation protocol (Frisoni and Jack 2011, Frisoni *et al* 2015), consensus has not been reached. Therefore, a segmentation algorithm with the ability to adapt to different protocols is very desirable.

Based on the previous considerations, in this paper we present a novel and fully automated hippocampal segmentation algorithm, named HUMAN (Hippocampal Unified Multi-Atlas-Networks), which combines in a unified framework the accuracy of multi-atlas methods with the robustness of artificial neural networks classification. The performance of this methodology was assessed with two independent test sets, segmented with two independent protocols. The first set was provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the second one by the Open Access Series of Imaging Studies (OASIS). Different manifold strategies for atlas selection were explored in order to identify an optimal setup for learning. The performance of HUMAN when applied to sub-regions (*patches*) of the hippocampus was then investigated. Finally, HUMAN was compared to the publicly available segmentation tool FreeSurfer (Fischl 2012) and to a basic multi-atlas pipeline, consisting of registration and label fusion.

2. Materials

A data set of 100 T1 MRI scans from the ADNI database (1.5 T and 3.0 T), including normal control (NC), mild cognitive impairment (MCI) and Alzheimer's disease (AD) subjects, was used in preparation of this article. The relative hippocampal labels were provided by the EADC-ADNI harmonized segmentation protocol[7] (Boccardi *et al* 2015a, 2015b). The ADNI images were divided in two data sets matched for size and demographic features. The first one $\mathcal{T}_{\text{ADNI}}$, consisting of 45 images, was used for training and parameter tuning. The second one $\mathcal{D}_{\text{ADNI}}$, of 55 images, was used as a test set. $\mathcal{T}_{\text{ADNI}}$ and $\mathcal{D}_{\text{ADNI}}$ shared common acquisition characteristics and the same harmonized segmentation protocol.

A further evaluation was performed on an independent $\mathcal{D}_{\text{OASIS}}$ set, consisting of 35 T1 MRI scans (1.5 T), provided by the OASIS initiative[8] (Marcus *et al* 2007) in occasion of the MICCAI SATA challenge workshop 2013[9] with the relative labels provided by the braincolor protocol[10] (Klein *et al* 2010). Both the ADNI and $\mathcal{D}_{\text{OASIS}}$ sets consisted of MPRAGE MRI brain scans with a resolution of $1 \times 1 \times 1$ mm^3(in the following paragraphs voxels and mm^3 are interchangeably used).

Data size, clinical status, age and gender information for the three sets $\mathcal{T}_{\text{ADNI}}$, $\mathcal{D}_{\text{ADNI}}$ and $\mathcal{D}_{\text{OASIS}}$ are summarized in table 1. Left and right hippocampal volume averages are reported with the relative standard deviations. The age range for $\mathcal{D}_{\text{OASIS}}$ is consistently broader then $\mathcal{D}_{\text{ADNI}}$ as the OASIS project was not limited to elderly subjects. However, this difference does not affect the reliability of the results.

Clinical and gender information for the $\mathcal{D}_{\text{OASIS}}$ set was not available. $\mathcal{T}_{\text{ADNI}}$ and $\mathcal{D}_{\text{ADNI}}$ were matched in terms of demographic and clinical composition. The volume distributions in the training set and test set were also matched, thus excluding any volume-based bias in the analysis. The image processing and the learning phases were carried out blindly to subject status.

[7] www.hippocampal-protocol.net
[8] www.oasis-brains.org
[9] https://masi.vuse.vanderbilt.edu/workshop2013
[10] www.braincolor.org

Table 1. Group size, age range, gender, hippocampal volumes (mean and standard deviation) and clinical composition (normal control NC, mild cognitive impairment MCI and Alzheimer's disease AD subjects) of the training and test data sets.

Data	Size	Age (years)	M/F	Right Vol. (mm^3)	Left Vol. (mm^3)	Subjects
$\mathcal{T}_{\text{ADNI}}$	45	60–90	24/21	3780 ± 660	3693 ± 634	15 NC–15 MCI–15 AD
$\mathcal{D}_{\text{ADNI}}$	55	63–88	32/23	3597 ± 578	3559 ± 593	14 NC–19 MCI–22 AD
$\mathcal{D}_{\text{OASIS}}$	35	18–90	N/A	3788 ± 436	3577 ± 426	N/A

Note: The groups are matched for hippocampal volume averages and relative standard deviations.

3. Methods

The rationale underlying the HUMAN approach is to emulate the manual segmentation of a human expert within a multi-atlas framework. It cannot be considered a machine learning segmentation method, as its goal is not the generalization of models learned from training examples, nor a label fusion strategy, as the core of the method is the generation of putative segmentations and not the fusion of propagated labels. The novel algorithm combines multi-atlas and classification approaches and involves three main phases:

- **Nonlinear registration**. MRI scans are intensity normalized and non-linearly registered with a data driven template. The goal of this processing step is to increase the similarity among the scans as far as possible. Volumes of Interest (VOIs) are extracted from each warped scan to define a peri-hippocampal region of interest.
- **Atlas selection**. The VOIs and the displacement fields resulting from non-linear registration are used to perform linear and non-linear similarity measurements between the test image and the training scans. Accordingly this step defines which atlases should be used as base of knowledge for subsequent learning and classification.
- **Classification**. VOIs of selected atlases undergo a feature extraction process, the resulting statistical and textural features are then used to train a voxel-based classifier for each VOI. A test VOI undergoes the same feature extraction process then the selected classifiers are used to estimate whether a voxel belongs or not to the hippocampus. The hippocampal segmentation in the test images is finally obtained by label fusion.

Figure 1 shows a synthetic overview of the algorithm. The full method is illustrated in the following and further methodological aspects are discussed in the supplementary material (stacks.iop.org/PMB/60/8851/mmedia).

3.1. Nonlinear registration

Since registration is sensitive to the initial conditions, the intensities of the brain scans were normalized and the bias field removed with the improved N3 MRI bias field correction algorithm (Tustison *et al* 2010). After pre-processing, one image a_v was repeatedly extracted from the $\mathcal{T}_{\text{ADNI}}$ set to perform a leave-one-out analysis. The healthy controls from the remaining training set $\mathcal{D}_t$ were used to build a data driven template $\mathcal{M}_t$ (see figure 2) to facilitate data registration using the advanced normalization tools[11] (ANTs) (Avants *et al* 2009, 2011). Leave-one-out was adopted for template construction in order to faithfully reproduce the segmentation process of test scans and maximize the computational efficiency of the method, not requiring a dedicated template for each test scan.

[11] http://picsl.upenn.edu/software/ants/

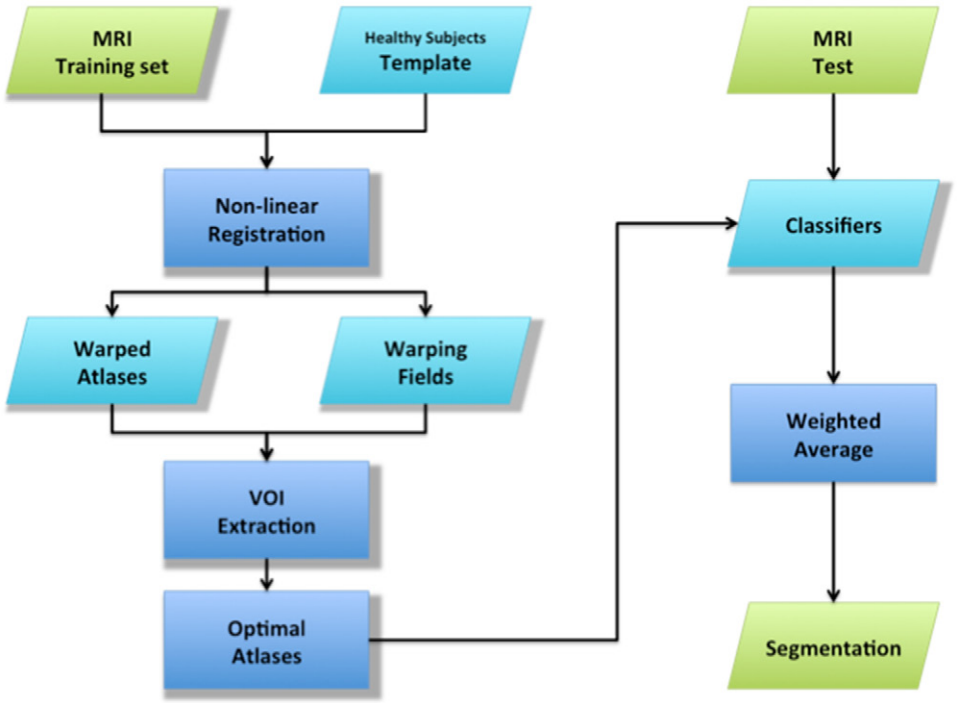


Figure 1. Synthetic overview of the proposed method. Healthy subjects of the training set are used to build a data driven template, then all training scans are non-linearly registered and hippocampal volumes of interest (VOI) extracted. Warped atlases and warping fields are stored for later use. A test MRI scan is warped to the template and the most similar examples are selected according to a similarity metric. Each optimal atlas is used to train a dedicated classifier and to obtain a putative segmentation. Finally, the test segmentation is obtained by averaging the putative segmentation according to the adopted similarity metric.

For each cross-validation round, the $\mathcal{D}_t$ brain scans and the validation image a_v were linearly registered to the $\mathcal{M}_t$ template with FSL-FLIRT (Jenkinson *et al* 2012). Then, a non-linear registration (Klein *et al* 2009) was performed with ANTs, and the warp fields $\mathcal{F}_i$ were stored for later use.

After registration a gross peri-hippocampal region $\omega(\text{VOI})_i$ and the corresponding field $\mathcal{F}(\text{VOI})_i$ were extracted, from both training and test, using FAPoD (Amoroso *et al* 2012, Tangaro *et al* 2014) (a fully automated hippocampal shape analysis algorithm). The $\omega(\text{VOI})_i$ contained a probable hippocampal region of about 17 000 voxels, and laid in a rectangular region of interest of dimensions $50 \times 70 \times 70$ mm^3. $\omega(\text{VOI})_i$ and the relative warp field $\mathcal{F}(\text{VOI})_i$ were used for the subsequent atlas selection.

3.2. Atlas selection

Two different strategies were adopted in order to select the optimal atlases. In the first strategy, as suggested by previous studies (Gerber *et al* 2010), we used embedding techniques to project the peri-hippocampal $\omega(\text{VOI})_i$ voxel intensities, and the related voxel-wise warp displacements $\mathcal{F}(\text{VOI})_i$ (accounting for $\sim$17 000 voxels), into low dimensional manifolds. Subsequently, the k atlases nearest to the volume of interest of the validation image $\omega(\text{VOI})_v$

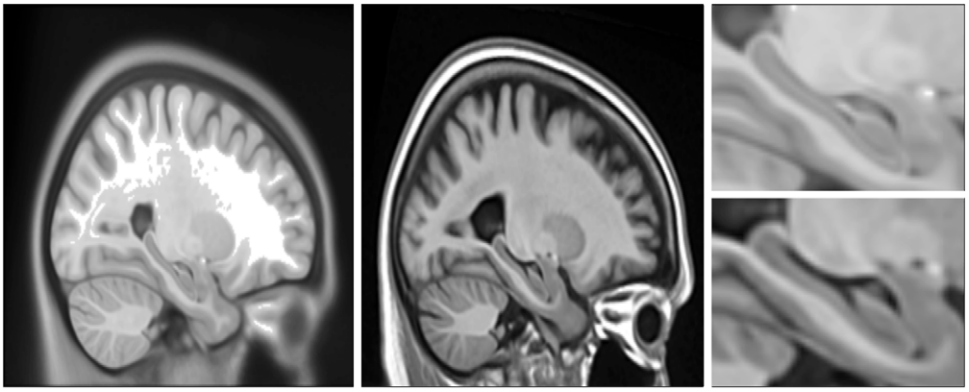


Figure 2. The MNI152 non-linear T1 weighted template (left) is qualitatively compared with a data-driven template $\mathcal{M}_t$ obtained by averaging only healthy subjects (middle) showing a sagittal slice containing the right hippocampus. The right panel shows (top) a magnified view of the MNI152 peri-hippocampal region and the corresponding region for the $\mathcal{M}_t$ template (bottom).

were selected as optimal. Several parametric and non-parametric techniques were explored for this task: Sammon mapping (SAM) (Sammon 1969), Isomap (ISO) (Tenenbaum *et al* 2000), locally linear embedding (LLE) (Roweis and Saul 2000), Laplacian Eigenmaps (LAP) (Belkin and Niyogi 2001), stochastic neighbor embedding (SNE) (Hinton and Roweis 2002) and its improved version (t-SNE) (Van der Maaten and Hinton 2008). The second strategy consisted in using the Pearson's correlation to measure directly the similarity among the peri-hippocampal regions $\omega(\text{VOI})_i$ and $\omega(\text{VOI})_v$.

For each dimensionality reduction technique different parameter configurations were considered, resulting in the selection of different atlases. In particular, for each manifold we explored the number of atlases to be selected, ranging from 1 to 30, and the embedding manifold dimension. This range was chosen based on the fact that multi-atlas performances usually degrade when using more than $15 \sim 20$ atlases (Aljabar *et al* 2009). The Dice similarity index (see section 3.3) was used to evaluate the leave-one-out best configuration.

3.3. Classification and segmentation

The hippocampal $\omega(\text{VOI})_i$ belonging to the k selected optimal atlases underwent a statistical and textural feature extraction process (Tangaro *et al* 2015). For each voxel Haralick, Haar-like and statistical features such as, average, standard deviation, kurtosis, skewness and gradients were computed. The relationships between each voxel and the voxels surrounding it were taken into account using varying size windows centered on the examined voxel with dimensions ranging from $3 \times 3 \times 3$ voxels to $9 \times 9 \times 9$ voxels, for a whole set of 315 features (Tangaro *et al* 2014), thus each scan was described as a matrix of approximate dimensions $17\,000 \times 315$.

Subsequently, a k-tuple of neural network classifiers $\mathcal{C}_{\{1,\ldots,k\}}$ was trained. Since the aim of this approach is to use warping to increase as much as possible the similarity between the test scan and the training images, we trained the networks $\mathcal{C}_i$ to exactly represent the corresponding training data $\omega(\text{VOI})_i$. We also investigated whether the classification performance was locally robust when training the $\mathcal{C}_{\{1,\ldots,k\}}$ on hippocampal sub-regions, called 'patches'. These patches were introduced by subdividing the $\omega(\text{VOI})_i$ regions in equally spaced $10 \times 10 \times 10$ voxel windows, thus obtaining $\alpha = 245$ patches. In this case, we trained the classifiers $\mathcal{C}^{\alpha}_{\{1,\ldots,k\}}$

for each α hippocampal local patch instead of considering the whole peri-hippocampal region $\omega(\mathrm{VOI})_i$, thus resulting in a geometrical pruning of the feature space. The classification performances were measured with the Dice similarity index:

$$\mathrm{Dice} = \frac{2\,|A \cap B|}{|A|+|B|}$$

with A and B representing the regions being compared and cardinalities $|A|, |B|$ intended as the measured volumes.

The best classification results were obtained with artificial neural networks, trained with the backpropagation algorithm, consisting of one hidden layer with ten neurons and standard sigmoid activation functions. With this design, the networks achieved Dice indexes ranging from 0.98 to 1.00. Networks trained with a lower number of neurons could not achieve such performances, while no significant improvement could be obtained with higher numbers of neurons. Each atlas was used simultaneously as a training and testing scan, in order to build a model of the atlas itself; then these trained models were used to generate putative segmentations of the test scans. The same configuration was maintained for the networks trained on the hippocampal patches. In the HUMAN approach the test images are processed to increase the similarity with the training atlases. Therefore, the training of the classifier was aimed to model the atlases rather than generalize to unseen data samples, this is why we chose to adopt a more versatile classifier (artificial neural networks) instead of a more robust classifier, such as Random Forests. The trained models were finally stored.

For each validation scan, the segmentation was obtained by propagating the putative segmentations, as obtained from each network, onto the native target image space through the displacement inverse field $\mathcal{F}_t^{-1}$ and finally fusing the putative labels. More in detail, for each voxel the relative label was calculated as a weighted average of the k predicted labels, the weight being the pairwise distance between the selected atlases and the target image.

Several studies have shown that majority voting strategies for label fusion can yield hippocampal volumes significantly minor than those obtained by manual segmentation (Sabuncu *et al* 2010, Khan *et al* 2011, Wang *et al* 2011). This is mainly caused by the monotonic decrease of signal to noise ratio when moving from inner to outer hippocampal regions and consequently by an unbalanced error rate in favor of false negatives. To overcome this systematic error we used an adaptive threshold in the voxel classification phase, based on the Bayes theorem. We used the probability assigned by FAPoD Tangaro *et al* (2014) to each voxel to belong or not to the hippocampus as *a priori* probabilities $P(H)$. For a two class problem, given the average training sensitivity S, the classifier probability to correctly label hippocampal voxel $P(h|H)$, and specificity s, the classifier probability to correctly label background voxel $P(\neg h|\neg H)$, the probability of a voxel to be assigned to the hippocampus is:

$$\begin{aligned} P(h) &= P(h|H) * P(H) + P(h|\neg H) * P(\neg H) \\ &= S * P(H) + (1-s) * (1 - P(H)) \end{aligned}$$

Following the Bayes theorem the *a posteriori* probability for a voxel to belong to the hippocampus when positively labeled $P(H|h)$ is given by:

$$\begin{aligned} P(H|h) &= \frac{P(h|H) * P(H)}{P(h)} \\ &= \frac{S * P(H)}{S * P(H) + (1-s) * (1 - P(H))} \end{aligned}$$

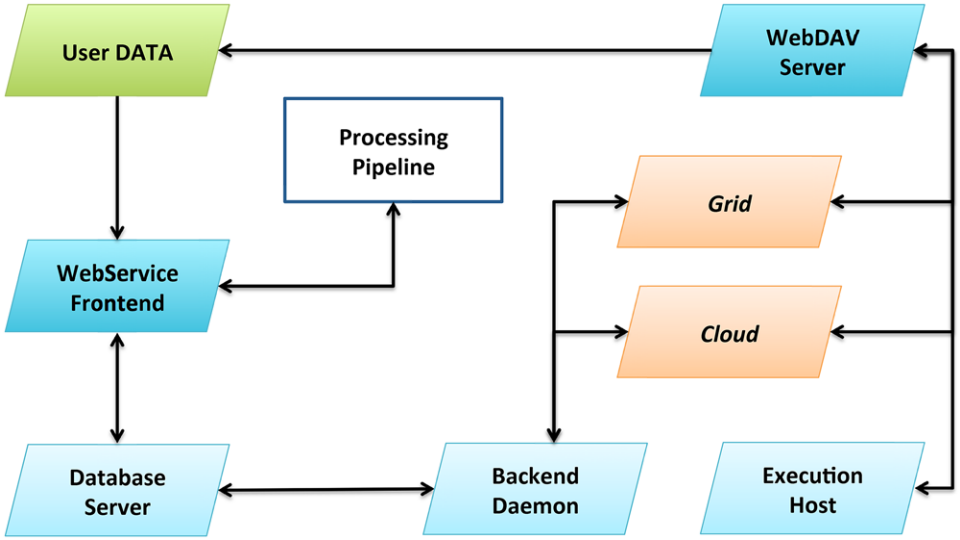


Figure 3. The user data are processed with a continuous Web Distributed Authoring and Versioning (WebDAV) protocol linked to the execution host. Different distributed infrastructures, such as computer grids or clouds, can be exploited according to user requirements.

This probability is then used as decision threshold. Accordingly, inner voxels which have higher *a priori* probability to belong to the hippocampus are assigned lower thresholds then outer ones, as a consequences the probability to have false negatives and the statistical error in the hippocampal volume evaluation are both reduced.

3.4. Computational infrastructure

This method requires a complex software framework, involving processing tools developed with different languages and in different environments. This could hinder the diffusion of its use in clinical or research settings lacking strong technological background. To overcome these challenges we developed a user friendly environment exposing Human as a Service, a schematic overview is presented in figure 3.

The computational resources for this study were provided by the ReCaS computer center (Bari, Italy)[12], a computing infrastructure, consisting of about 5000 CPU and allowing up to 2.2 PB storage. The data processing and monitoring was performed by exploiting a dynamic job submission tool (JST) facility (Amoroso *et al* 2014). JST is a job management tool particularly useful to manage the submission and monitoring of applications, when a large number of independent executions are needed to solve the required tasks. Different distributed infrastructures are suitable for HUMAN analyses, such as computer grids or clouds.

Hence this approach is flexible and easy to be implemented on dedicated destination machines. Moreover, the necessary software can be easily interfaced with web portals or common workflow manager tools. The HUMAN pipeline is also available, like a web service/cloud solution at the following link: https://recasgateway.ba.infn.it.

[12] www.recas-pon.ba.infn.it

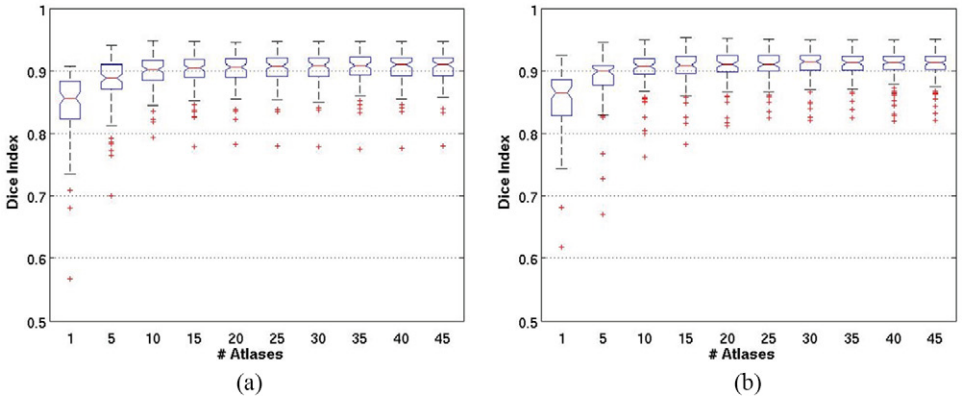


Figure 4. The figure shows how the Dice index performances vary with number of atlases for both left (a) and right (b) hippocampi. Best performances were achieved with about 10 atlases.

4. Results

4.1. Results for atlas selection

To identify the best atlas selection method the embedding techniques listed in section 3.2 were compared to a Pearson's correlation with a leave-one-out analysis. The best results, in terms of Dice index, were achieved by HUMAN using the Pearson's correlation between the training and the validation $\omega(\mathrm{VOI})_i$. For the left hippocampi, the median Dice index was 0.910 ± 0.004, and for the right hippocampi 0.914 ± 0.004. The analysis was performed using the optimal configuration found with a basic multi-atlas approach, previously defined as a multi-atlas consisting of just registration and label fusion.

The performance of the basic-multi atlas was never as good as that of HUMAN. The best results, obtained in this case with Pearson's correlation, indicated for the left hippocampi a median Dice index 0.869 ± 0.006 and for the right hippocampi 0.873 ± 0.005. Once established that the best method for atlas selection was the use of Pearson's correlation for the similarity measurements, we explored how the number of selected atlases would affect the segmentation performance.

In figure 4, the Dice index is represented as a function of the number of atlases. The best outcome was achieved with $\sim$10 atlases, after which a plateau was reached, with no significant difference (Wilcoxon $p > .05$). As a consequence further analyses were carried out considering only the best ten atlases.

4.2. Segmentation results for ADNI scans

The $\mathcal{D}_{\mathrm{ADNI}}$ test set shared appearance features and segmentation protocol with the training set $\mathcal{T}_{\mathrm{ADNI}}$. Moreover, they were matched in terms of demographic and clinical composition. To assess the method performances and the relative segmentation quality, a Bland–Altman analysis (Bland and Altman 1995) was performed (see figure 5) along with the Dice index measurements.

Segmented volumes and manual tracings showed a very high correlation (0.95 and 0.96 for respectively left and right hippocampi). The 95% confidence interval limits were almost the same for both left and right hippocampi $[-400, 400]$. These values along with the high correlation suggest good agreement between segmented and manual volumes. Moreover we evaluated the manual versus

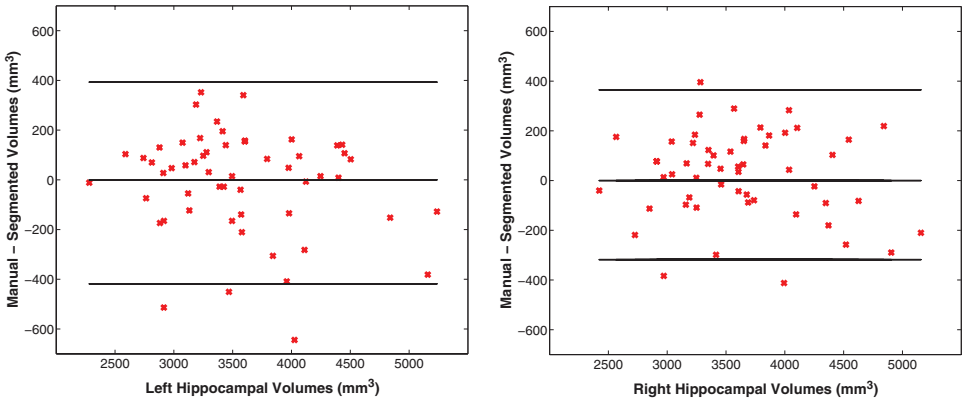


Figure 5. The figure shows the $\mathcal{D}_{\text{ADNI}}$ results obtained for left (a) and right (b) hippocampi. The Bland–Altman plots show the agreement between the manual tracings and the automated segmentations within the 95% confidence level.

segmented volume difference distribution obtaining a mean value of -12.6 ± 207 voxels, which did not significantly differ from zero. It is evident that no significant bias affected the segmented volumes. The segmentation agreement was also measured in terms of Dice index, 0.926 ± 0.003 and 0.931 ± 0.002 for left and right hippocampi respectively (0.929 ± 0.003 on average).

The performance on the test set $\mathcal{D}_{\text{ADNI}}$ was also compared with the multi-atlas segmentation procedure described in section 4.1 and FreeSurfer segmentations. Also in this case the optimal configuration determined in training was adopted. The results of this comparison are presented in the following table 2.

4.3. The segmentation protocol effect on MICCAI scans

The performance of HUMAN was evaluated with the protocol independent test set $\mathcal{D}_{\text{OASIS}}$ provided, as previously mentioned, by OASIS in occasion of the MICCAI SATA challenge workshop 2013. As described in section 4.2 the method was assessed with a correlation and a Bland–Altman analysis. The results for both left and right segmentations are shown in figure 6.

The correlation between the volumes segmented with HUMAN and those segmented manually was high (left correlation is 0.83, right 0.79) even if lower than in the former case. Bland–Altman analysis showed a quite broad 95% confidence interval with similar values for both hippocampi $\sim[-500,500]$, nevertheless satisfactory levels of accuracy in terms of median Dice index were achieved (0.856 ± 0.002 and 0.862 ± 0.002 respectively for left and right hippocampi; on average 0.869 ± 0.002). For the left difference distribution (manual—segmented volumes) we found on average -82 ± 222 voxels, for right hippocampi -26 ± 296; for both cases no significant bias was detected.

HUMAN performances were also significantly better than those achieved by multi-atlas and FreeSurfer. A summary of the results is presented in table 3.

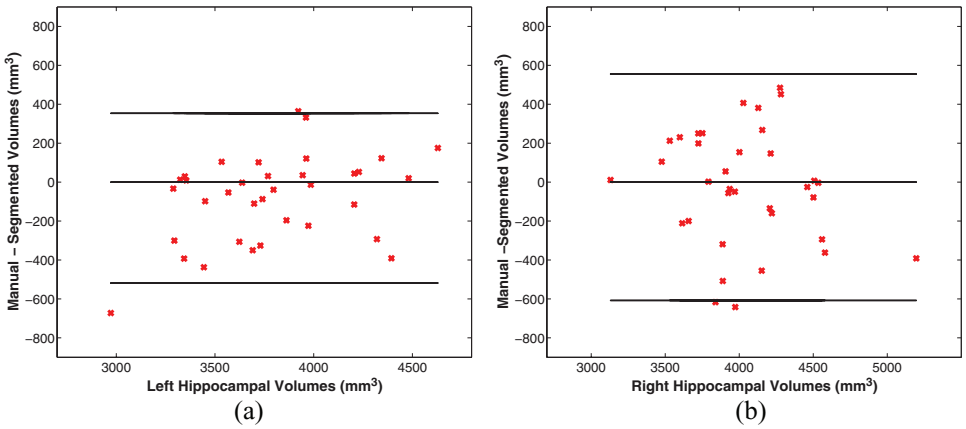
4.4. HUMAN scale robustness

In the final phase of this work, we explored the performance of HUMAN within a patch-based segmentation framework. In this case, the goals of the analysis were twofold: to investigate

Table 2. Median Dice indexes and corresponding standard errors for HUMAN, Multi-atlas and FreeSurfer segmentations.

	HUMAN	Multi atlas	FreeSurfer
Left hippocampus	0.926 ± 0.003	0.861 ± 0.006	0.707 ± 0.006
Right hippocampus	0.931 ± 0.002	0.869 ± 0.004	0.715 ± 0.006
Left volumes (mm^3)	3562 ± 664	3528 ± 420	3337 ± 752
Right volumes (mm^3)	3640 ± 631	3610 ± 458	3320 ± 557

Note: The segmented volumes for both left and right hippocampi are also shown.



(a) (b)

Figure 6. The figure shows the $\mathcal{D}_{\text{OASIS}}$ results obtained for left (a) and right (b) hippocampi. For both cases a 95% level confidence agreement is shown through the Bland–Altman plots.

Table 3. Median Dice indexes and standard errors for HUMAN, Multi-atlas and FreeSurfer, and the segmented volumes for both left and right hippocampi (average and standard deviation of the distribution).

	HUMAN	Multi atlas	FreeSurfer
Left hippocampus	0.856 ± 0.002	0.825 ± 0.003	0.797 ± 0.005
Right hippocampus	0.862 ± 0.002	0.832 ± 0.003	0.808 ± 0.004
Left volumes (mm^3)	3834 ± 368	3414 ± 435	4200 ± 448
Right volumes (mm^3)	4024 ± 397	3530 ± 453	4150 ± 409

whether the classification would be affected by local hippocampal shape effects, and to investigate whether HUMAN performances would be uniformly distributed over the whole hippocampal shape.

This involved segmenting each α test patch with the most correlated $\mathcal{C}^{\alpha}_{\{1,\ldots,k\}}$ models. Also in this case, the final prediction was obtained by averaging the scores obtained by the k optimal classifiers, i.e. those trained on the patches better correlated with the patch to be segmented. The final segmentation was obtained by merging the patch segmentations.

The results confirmed the robustness of HUMAN throughout the whole hippocampus for both $\mathcal{D}_{\text{ADNI}}$ and $\mathcal{D}_{\text{OASIS}}$. This is illustrated in figure 7, where each patch is color-coded according to the relative dice obtained by averaging the $\mathcal{D}_{\text{ADNI}}$ and $\mathcal{D}_{\text{OASIS}}$ patch-based results.

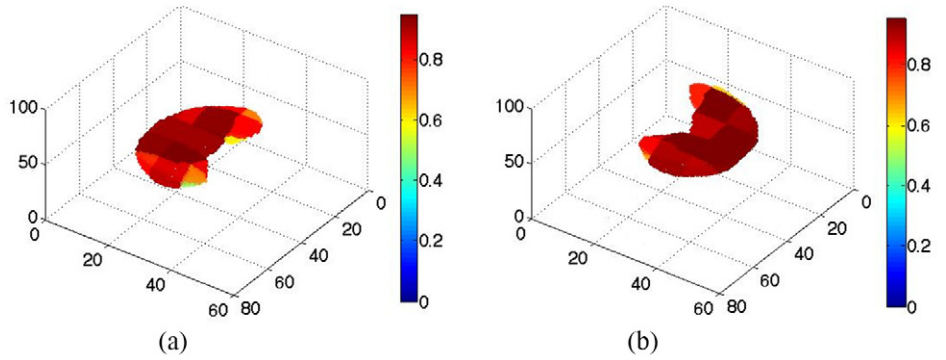


Figure 7. The figure shows the patch wise Dice distribution for both left (a) and right (b) hippocampi. The color-bars show the median Dice index values. Only few patches, related to hippocampal head and tail, show dice indexes smaller than 0.7.

In particular, for $\mathcal{D}_{\text{ADNI}}$ we were able to obtain a small but significant ($p < .05$ for both left and right hippocampi) improvement. The Dice index increased from 0.905 ± 0.004 to 0.910 ± 0.003 and from 0.913 ± 0.003 to 0.922 ± 0.003, respectively for left and right hippocampi. For $\mathcal{D}_{\text{OASIS}}$ the Dice index remained constant for the left hippocampi (0.847 ± 0.003) and increased from 0.852 ± 0.003 to 0.857 ± 0.003 ($p < .05$) for the right hippocampi. Significance was assessed also in this case with a Wilcoxon test.

The poorest results were obtained for patches situated at the head of the hippocampus. Figure 8 shows a qualitative comparison among segmentations for 12 randomly sampled subjects (from $\mathcal{D}_{\text{ADNI}}$): 4 NC, 4 MCI and 4 AD subjects.

5. Discussion and conclusion

In this study we presented a novel segmentation algorithm—HUMAN—based on a combined multi-atlas and machine learning strategy. HUMAN produced accurate segmentation on two independent test sets. In the first test set, $\mathcal{D}_{\text{ADNI}}$, with manual labels traced with the same protocol of $\mathcal{T}_{\text{ADNI}}$, the segmentation results were excellent with median Dice index $= 0.929 \pm 0.003$ for both left and right hippocampi. In the second test set $\mathcal{D}_{\text{OASIS}}$, with tracings obtained with a different protocol, the performance of HUMAN (Dice index 0.869 ± 0.002) were less impressive, but yet satisfactory if compared with other recently reported studies (Cardoso *et al* 2013, Kim *et al* 2013, Kwak *et al* 2013, Pipitone *et al* 2014). While the best performing methods of MICCAI SATA challenge reported Dice indexes approaching 0.90 median values, based on its performance on $\mathcal{D}_{\text{OASIS}}$, HUMAN would have still placed itself among the best five performing algorithms[13]. Besides, one should take into account that the performances reported (Iglesias *et al* 2012, Wang and Yushkevich 2013, Zikic *et al* 2013) were achieved with training and test sets sharing the same segmentation protocol, while HUMAN was trained with a different hippocampal segmentation protocol.

In addition, we tried to tackle the following questions: (i) can machine learning strategies bring a substantial improvement to state-of-the-art segmentation strategies, such as the multi-atlas approaches? (ii) to which degree are machine learning strategies affected by the use of

[13] http://masi.vuse.vanderbilt.edu/submission/leaderboard_final.html

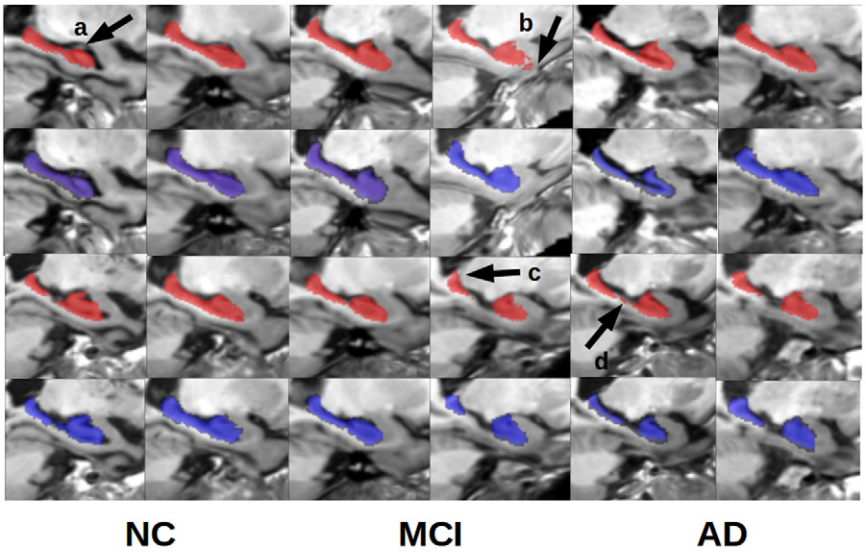


Figure 8. Human segmentations (first row and third row) with the corresponding manual tracings (second row and fourth row) of 4 randomly chosen examples for controls (NC), mild cognitive impairment subjects (MCI) and Alzheimer's disease subjects (AD). Some segmentation errors are also underscored for the hippocampal head ((a), (b)), tail (c) and body (d).

different segmentation protocols? (iii) how much are machine learning strategies affected by local, or patch based, training?

We demonstrated that multi-atlas could be significantly improved if combined with a machine learning strategy. Moreover this improvement was robust to segmentation protocol differences between training and test. In fact, for both test sets $\mathcal{D}_{\text{ADNI}}$ and $\mathcal{D}_{\text{OASIS}}$, a significant improvement (of about 5.5%) was found. The differences in protocol, as expected, affected the method accuracy. In particular, this could be observed with a loss of correlation between manual and HUMAN segmentations. As expected FreeSurfer performances resulted lower than those achieved by HUMAN, however FreeSurfer is trained with a different segmentation protocol and so this comparison biased against it. Nonetheless, it is interesting to note that HUMAN performances showed a greater stability with a variation of about 6.4% against the 11.4% of FreeSurfer. Finally, the use of local training (patch based) demonstrated the robustness of the method when dealing with sub-hippocampal regions. With this latter patch-based method the segmentation performances resulted slightly improved and almost uniformly distributed over the hippocampus. The results also confirmed the heads of the hippocampus are the regions presenting more difficulties in terms of segmentation.

In our previous work (Tangaro *et al* 2014, Inglese *et al* 2015) we discussed an entirely machine learning based segmentation procedure. In particular, we used a peri-hippocampal region to actively determine a set of training images, which were then used to train a unique classifier. In this work we propose a complete change of paradigm. HUMAN exploits intensity and spatial normalization techniques to best fit the test data to the training set and to define an optimal base of knowledge (putative segmentations) to be combined in a multi-atlas framework.

As previously remarked, other recent work has already shown interesting results on this approach (Hao *et al* 2014). However, the fundamental difference between the method here described and the above mentioned combined label fusion strategies, is that they introduce

voxel-wise machine learning strategies for label fusion, while our method uses machine learning to determine new putative atlases, obtaining the label fusion through a weighted majority voting procedure.

The slight performance deterioration on $\mathcal{D}_{\text{OASIS}}$ confirmed that segmentation protocols play a key-role when it comes to multi-center studies. To the best of our knowledge, this was the first study directly addressing the effects of two independent segmentation protocols on fully automated segmentation techniques.

It is worthwhile to note that the proposed method is fully automated (it does not require user intervention) and computationally efficient, requiring a processing time of about 10 min per test image. The possible exploitation of cloud infrastructures implies that it could be adopted for large clinical trials. A limitation of the study lies in the absence of clinical evaluation, which was outside the goals of this work. Recent literature has shown that hippocampal sub-regions could be important as quantitative bio-markers for a number of neurodegenerative diseases, and especially Alzheimer's disease. Bland–Altman plots, Dice index and correlation measurements all confirm that HUMAN segmentation are consistent with the manual tracings. The adoption of a Bayesian strategy for adaptive thresholding has consistently improved the segmentation performance, nonetheless a further improvement of the presented method could consider the exploration of more refined label fusion strategies. Another refinement of the method could eventually investigate which features contribute the most to an optimal label combination.

In terms of future developments, we plan to apply HUMAN to a clinical data set with the aim of assessing its validity as a tool aiding the diagnosis of Alzheimer's disease, as we did in our previous work on pattern recognition (Amoroso *et al* 2014, Sensi *et al* 2014, Bron *et al* 2015). Both the high Dice index and correlation values, especially obtained for $\mathcal{D}_{\text{ADNI}}$ scans, would suggest HUMAN applicability to both studies based on changes of volume due to disease progression and group-wise comparisons with fixed reference volumes, for example for MCI-AD transition.

Acknowledgments

N Amoroso, R Errico and A Tateo acknowledge funding by the Italian MIUR grant PON PRISMA Cod. PON04a2_A. This research was also supported by Istituto Nazionale di Fisica Nucleare (INFN), Italy.

Data used in the preparation of this article was obtained from the 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 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 M 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. 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.

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

Aljabar P, Heckemann R A, Hammers A, Hajnal J V and Rueckert D 2009 Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy *NeuroImage* **46** 726–38

Amoroso N, Bellotti R, Bruno S, Chincarini A, Logroscino G, Tangaro S and Tateo A 2012 Automated shape analysis landmarks detection for medical image processing *Proc. of the Int. Symp., CompIMAGE* pp 139–42

Amoroso N, Errico R and Bellotti R 2014 PRISMA-CAD: fully automated method for computer-aided diagnosis of dementia based on structural MRI data *Proc. of the Computer-Aided Diagnosis of Dementia Based on Structural MRI Data, MICCAI* pp 16–24

Amoroso N, Errico R, Ferraro G, Tangaro S, Tateo A and Bellotti R 2014 Fully automated MRI analysis for brain diseases with high performance computing *SCORE@POLIBA* pp 347–51

Avants B B, Tustison N J, Song G, Cook P A, Klein A and Gee J C 2011 A reproducible evaluation of ants similarity metric performance in brain image registration *NeuroImage* **54** 2033–44

Avants B B, Tustison N and Song G 2009 Advanced normalization tools ants *Insight J.* **2** 1–35

Belkin M and Niyogi P 2001 Laplacian eigenmaps and spectral techniques for embedding and clustering *15th Annual Neural Information Processing Systems Conf.* vol 14 pp 585–91

Bellotti R and Pascazio S 2012 Editorial: advanced physical methods in brain research *Eur. Phys. J. Plus* **127** 145

Bland J M and Altman D G 1995 Comparing methods of measurement: why plotting difference against standard method is misleading *Lancet* **346** 1085–7

Boccardi M *et al* 2015a Delphi definition of the EADC-ADNI Harmonized protocol for hippocampal segmentation on magnetic resonance *Alzheimer's Dementia* **11** 126–38
Boccardi M *et al* 2015b Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol *Alzheimer's Dementia* **11** 183–91
Bron E E *et al* 2015 Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the caddementia challenge *NeuroImage* **111** 562–79
Bruno S D, Cercignani M and Wheeler-Kingshott C A M 2012 Neurodegenerative dementias: from MR physics lab to assessment room *Eur. Phys. J. Plus* **127** 1–15
Cardoso M J, Leung K, Modat M, Keihaninejad S, Cash D, Barnes J, Fox N C and Ourselin S 2013 Steps: similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation *Med. Image Anal.* **17** 671–84
Chincarini A *et al* 2013 Automatic temporal lobe atrophy assessment in prodromal ad: data from the descripa study *Alzheimers Dementia* **1** 12
Duc A K H *et al* 2013 Using manifold learning for atlas selection in multi-atlas segmentation *PloS One* **8** e70059
Fischl B 2012 FreeSurfer *NeuroImage* **62** 774–81
Frisoni G B *et al* 2015 The EADC-ADNI harmonized protocol for manual hippocampal segmentation on magnetic resonance: evidence of validity *Alzheimers Dementia* **11** 111–125
Frisoni G B and Jack C R 2011 Harmonization of magnetic resonance-based manual hippocampal segmentation: a mandatory step for wide clinical use *Alzheimers Dementia* **7** 171–4
Gerber S, Tasdizen T, Thomas Fletcher P, Joshi S and Whitaker R 2010 Manifold modeling for brain population analysis *Med. Image Anal.* **14** 643–53
Hao Y, Wang T, Zhang X, Duan Y, Yu C, Jiang T and Fan Y 2014 Local label learning (l1l) for subcortical structure segmentation: application to hippocampus segmentation *Human Brain Mapp.* **35** 2674–97
Heckemann R A *et al* 2010 Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation *Neuroimage* **51** 221–7
Hinton G E and Roweis S T 2002 Stochastic neighbor embedding *Adv. Neural Inf. Process. Syst.* pp 833–40 (www.cs.nyu.edu/~roweis/papers/sne_final.pdf)
Iglesias J E, Sabuncu M R and Van Leemput K 2012 A generative model for probabilistic label fusion of multimodal data *Proc. of the Second Int. Conf. Multimodal Brain Image Anal.* pp 115–33
Inglese P *et al* 2015 Multiple RF classifier for the hippocampus segmentation: method and validation on EADC-ADNI harmonized hippocampal protocol *Physica Medica* at press
Jenkinson M, Beckmann C F, Behrens T E, Woolrich M W and Smith S M 2012 FSL *NeuroImage* **62** 782–90
Khan A R, Cherbuin N, Wen W, Anstey K J, Sachdev P and Beg M F 2011 Optimal weights for local multi-atlas fusion using supervised learning and dynamic information superdyn: validation on hippocampus segmentation *NeuroImage* **56** 126–39
Kim M, Wu G, Li W, Wang L, Son Y D, Cho Z H and Shen D 2013 Automatic hippocampus segmentation of 7.0 Tesla MR images by combining multiple atlases and auto-context models *NeuroImage* **83** 335–45
Klein A *et al* 2009 Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration *NeuroImage* **46** 786–802
Klein A, Dal Canton T, Ghosh S S, Landman B, Lee J and Worth A 2010 Open labels: online feedback for a public resource of manually labeled brain images *16th Annual Meeting for the Organization of Human Brain Mapping*
Klein S, van der Heide U A, Lips I M, van Vulpen M, Staring M and Pluim J P 2008 Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information *Med. phys.* **35** 1407–17
Kwak K, Yoon U, Lee D K, Kim G H, Seo S W, Na D L, Shim H J and Lee J M 2013 Fully-automated approach to hippocampus segmentation using a graph-cuts algorithm combined with atlas-based segmentation and morphological opening *Magn. Reson. Imaging* **31** 1190–6
Leung K K, Barnes J, Ridgway G R, Bartlett J W, Clarkson M J, Macdonald K, Schuff N, Fox N C and Ourselin S 2010 Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease *NeuroImage* **51** 1345–59
Lötjönen J M *et al* 2010 Fast and robust multi-atlas segmentation of brain magnetic resonance images *Neuroimage* **49** 2352–65

Maglietta R, Amoroso N, Boccardi M, Bruno S, Chincarini A, Frisoni G B, Inglese P, Redofi A, Tangaro S, Tateo A and Bellotti R 2015 Automated hippocampal segmentation in 3D MRI using random undersampling with boosting algorithm *Pattern Anal. Applic.* (doi:10.1007/s10044-015-0492-0)

Marcus D S, Wang T H, Parker J, Csernansky J G, Morris J C and Buckner R L 2007 Open access series of imaging studies oasis: cross-sectional MRI data in young, middle aged, nondemented, and demented older adults *J. Cogn. Neurosci.* **19** 1498–507

Morra J H, Tu Z, Apostolova L G, Green A E, Toga A W and Thompson P M 2010 Comparison of adaboost and support vector machines for detecting alzheimer's disease through automated hippocampal segmentation *IEEE Trans. Med. Imaging* **29** 30–43

Nestor S M *et al* 2013 A direct morphometric comparison of five labeling protocols for multi-atlas driven automatic segmentation of the hippocampus in Alzheimer's disease *NeuroImage* **66** 50–70

Pipitone J *et al* 2014 Multi-atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates *Neuroimage* **101** 494–512

Roweis S T and Saul L K 2000 Nonlinear dimensionality reduction by locally linear embedding *Science* **290** 2323–6

Sabuncu M R *et al* 2011 The dynamics of cortical and hippocampal atrophy in Alzheimer disease *Arch. Neurology* **68** 1040–8

Sabuncu M R, Yeo B T, Van Leemput K, Fischl B and Golland P 2010 A generative model for image segmentation based on label fusion *IEEE Trans. Med. Imaging* **29** 1714–29

Sammon J W 1969 A nonlinear mapping for data structure analysis *IEEE Trans. Comput.* **18** 401–9

Sensi F, Rei L, Gemme G, Bosco P, Amoroso N and Chincarini A 2014 Gdi*, a novel tool for mtl atrophy assessment *Proc. of the Computer-Aided Diagnosis of Dementia Based on Structural MRI Data, MICCAI* pp 92–101

Tangaro S *et al* 2014 Automated voxel-by-voxel tissue classification for hippocampal segmentation: methods and validation *Phys. Med.* **30** 878–87

Tangaro S *et al* 2015 Feature selection based on machine learning in mris for hippocampal segmentation *Comput. Math. Methods Med.* (online)

Tangaro S, Amoroso N, Bruno S, Chincarini A, Frisoni G B, Maglietta R, Tateo A and Bellotti R 2013 Active learning machines for automatic segmentation of hippocampus in MRI *Industrial Conf. on Data Mining, LNCS*

Tenenbaum J B, De Silva V and Langford J C 2000 A global geometric framework for nonlinear dimensionality reduction *Science* **290** 2319–23

Tustison N J, Avants B B, Cook P A, Zheng Y, Egan A, Yushkevich P A and Gee J C 2010 N4itk: improved n3 bias correction *IEEE Trans. Med. Imaging* **29** 1310–20

Van der Maaten L and Hinton G 2008 Visualizing data using t-SNE *J. Mach. Learn. Res.* **9** 2579–605

Wang H *et al* 2011 A learning-based wrapper method to correct systematic errors in automatic image segmentation: consistently improved performance in hippocampus, cortex and brain segmentation *NeuroImage* **55** 968–85

Wang H and Yushkevich P A 2013 Multi-atlas segmentation with joint label fusion and corrective learning an open source implementation *Front. Neuroinform.* **7** 27

Wolz R, Aljabar P, Hajnal J V, Hammers A and Rueckert D 2010 LEAP: learning embeddings for atlas propagation *NeuroImage* **49** 1316–25

Zikic D, Glocker B and Criminisi A 2013 Atlas encoding by randomized forests for efficient label propagation *Proc. Med. Image Comput. Comput.-Assist. Intervention* pp 66–73