Medical Image Analysis 17 (2013) 671–684

Contents lists available at SciVerse ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

STEPS: Similarity and Truth Estimation for Propagated Segmentations and its application to hippocampal segmentation and brain parcelation

M. Jorge Cardoso [a,*], Kelvin Leung [b], Marc Modat [a], Shiva Keihaninejad [b], David Cash [b], Josephine Barnes [b], Nick C. Fox [b], Sebastien Ourselin [a,b], for the Alzheimer's Disease Neuroimaging Initiative [1]

[a] Centre for Medical Image Computing (CMIC), University College London, UK
[b] Dementia Research Centre (DRC), University College London, UK

ARTICLE INFO

Article history:
Received 4 May 2012
Received in revised form 6 February 2013
Accepted 18 February 2013
Available online 1 March 2013

Keywords:
Label propagation
Local similarity metric
Hippocampus segmentation
Brain parcelation

ABSTRACT

Anatomical segmentation of structures of interest is critical to quantitative analysis in medical imaging. Several automated multi-atlas based segmentation propagation methods that utilise manual delineations from multiple templates appear promising. However, high levels of accuracy and reliability are needed for use in diagnosis or in clinical trials. We propose a new local ranking strategy for template selection based on the locally normalised cross correlation (LNCC) and an extension to the classical STAPLE algorithm by Warfield et al. (2004), which we refer to as STEPS for Similarity and Truth Estimation for Propagated Segmentations. It addresses the well-known problems of local vs. global image matching and the bias introduced in the performance estimation due to structure size. We assessed the method on hippocampal segmentation using a leave-one-out cross validation with optimised model parameters; STEPS achieved a mean Dice score of 0.925 when compared with manual segmentation. This was significantly better in terms of segmentation accuracy when compared to other state-of-the-art fusion techniques. Furthermore, due to the finer anatomical scale, STEPS also obtains more accurate segmentations even when using only a third of the templates, reducing the dependence on large template databases. Using a subset of Alzheimer's Disease Neuroimaging Initiative (ADNI) scans from different MRI imaging systems and protocols, STEPS yielded similarly accurate segmentations (Dice = 0.903). A cross-sectional and longitudinal hippocampal volumetric study was performed on the ADNI database. Mean ± SD hippocampal volume (mm^3) was 5195 ± 656 for controls; 4786 ± 781 for MCI; and 4427 ± 903 for Alzheimer's disease patients and hippocampal atrophy rates (%/year) of 1.09 ± 3.0, 2.74 ± 3.5 and 4.04 ± 3.6 respectively. Statistically significant ($p < 10^{-3}$) differences were found between disease groups for both hippocampal volume and volume change rates. Finally, STEPS was also applied in a multi-label segmentation propagation scenario using a leave-one-out cross validation, in order to parcellate 83 separate structures of the brain. Comparisons of STEPS with state-of-the-art multi-label fusion algorithms showed statistically significant segmentation accuracy improvements ($p < 10^{-4}$) in several key structures.

1. Introduction

The hippocampus, along with other structures in the medial temporal lobe, is one of the first structures where Alzheimer's disease pathology is earliest exhibited. It has been shown that measurements of hippocampal volume from volumetric T1 weighted MRI sequences and changes in volume from serial MRI sequences can aid in determining which subjects with mild cognitive impairment (MCI) will go on to develop Alzheimer's Disease (AD) (Jack et al., 1999; Ridha et al., 2007; Morra et al., 2009; Schuff et al., 2009; Henneman et al., 2009; Leung et al., 2010; Wolz et al., 2010) as well as predict cognitive decline in the earliest stages of the disease (Jack et al., 2000; Schuff et al., 2009; Wolz et al., 2010). As a result, there have been recent efforts to define the criteria for the presymptomatic (Sperling et al., 2011) and prodromal (Dubois et al., 2010) stages of Alzheimer's disease by incorporating biomarkers including hippocampal atrophy assessed using MRI. In a related effort, Jack et al. (2011) has recently provided guidelines for standardising and qualifying hippocampal volumetry and volume change measurements as a biomarker for use within clinical trials. One of the primary applications in clinical trials would be to use hippocampal volume as a criteria for eligibility into studies

* Corresponding author.
E-mail address: manuel.cardoso@ucl.ac.uk (M. Jorge Cardoso).

[1] Data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.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.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

1361-8415/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.media.2013.02.006

of therapeutic agents for patients with suspected prodromal AD and this has recently been recognised by regulatory authorities (EMA/CHMP/SWAP/809208/2011). The size of such studies mean that for quick targeted recruitment, efficient and accurate hippocampal volumetry techniques are needed. Manual methods are very time consuming (e.g. 45 min per hippocampus), require careful training and are demanding on the rater and therefore very expensive (Barnes et al., 2009; Frisoni and Jack, 2011). As a result, an automated hippocampal segmentation method that was accurate, robust and fast would be extremely valuable.

Segmentation propagation techniques make use of registration algorithms to align a manually labeled atlas to a new unsegmented image. The accuracy and robustness of this segmentation can be greatly improved by combining multiple candidate segmentations from a library/database of atlases (Rohlfing et al., 2004). In this way, the segmentation is dependent on the particular morphology of a single atlas and less vulnerable to errors in one or more regional labels. Each image from the atlas database, when registered to the image of interest, can be considered as an independent classifier. Several techniques for classifier fusion have been developed, where the most conventional method is a voting scheme (Xu et al., 1992).

As some registrations will be more accurate than others, more sophisticated techniques to quantify classifier performance have been developed. The classifiers can be combined according to their performance on a training set (Lam and Suen, 1995), or by estimating its performance on a feature or metric space (Woods et al., 1997). Aljabar et al. proposed to use the global normalised cross correlation between the registered template and the target image as a performance estimator in order to select the optimal classifiers for the voting scheme. While a global metric might be sufficient for simple shapes, the size of the database has to increase dramatically for objects with complex morphology in order to characterise the population's morphologic variability. Artaechevarria et al. (2009) proposed a set of local and global performance estimators based on image similarity metrics like the global normalised cross correlation (GNCC), global mean square difference (GMSD), global mutual information (GMI) and the corresponding local versions of the metrics, LNCC, LMSD and LMI respectively. Yushkevich et al. (2010) suggested a modified version of the LNCC metric using a ranking scheme and Collins and Pruessner (2010) used a GMI metric combined with a registration to a group-wise space in order to reduce computational cost. More recently, Sabuncu et al. (2010) reformulated the label fusion problem in a generative framework, providing a comprehensive probabilistic framework that rigorously motivates label fusion as a segmentation approach, by combining intensity similarity and a log-odds atlas propagation in a generative framework.

Instead of using an image similarity to derive the classifier performance, Warfield et al. (2004) proposed an algorithm named Simultaneous Truth and Performance Level Estimation (STAPLE) as a novel way to estimate the performance parameters of a classifier and consequently obtain the most likely classification. This framework estimates the classifier performance parameters by comparing each classifier to a consensus, in an iterative manner. It is important to note that the STAPLE framework was created for the purpose of fusing several manual or automated segmentations of the same image and not for fusing propagated segmentations. More recently, Asman and Landman (2012) and Commowick et al. (2012) introduced two reformulations of STAPLE with a spatially varying rater performance model that attempts to model miss registrations as part of the rater performance. However, these methodologies still find a local morphological consensus between the labels, without actually assessing the quality of the registration. Thus, the STAPLE framework and it's more recent reformulations do not explicitly incorporate the concept of atlas similarity or registration accuracy into the fusion model.

In summary, weighted voting techniques model segmentation errors as inaccuracies in the registration procedure and assume that the original manual segmentations do not have any mistakes, i.e. they are a ground truth. On the other hand, the STAPLE approach models segmentation errors as a rater performance problem instead of estimating registration accuracy and morphological similarity.

In order to make the STAPLE framework aware of registration errors, Leung et al. (2010) introduced the ranking concept used in Aljabar et al. (2009) into the STAPLE framework and showed improved segmentation accuracy. However, this global metric still suffers from the problems described above (e.g. complex morphology, local matching). Also, the GNCC metric is dependent on the ROI where it is calculated and is not robust to intensity non uniformity (INU) in MRI images.

We propose and validate a new strategy that incorporates a local similarity metric to estimate the expected image-based performance of each classifier on a voxel-by-voxel basis into a STAPLE formulation. This is the first time a local ranking and sampling strategy has been introduced into the STAPLE framework. We also introduce a new Markov Random Field (MRF) model optimised iteratively over the probabilistic labels in order to add spatial consistency and smoothness between the best local classifiers. This LNCC metric can cope with spatially variant registration accuracy, enabling the use of smaller template databases. Due to the local nature of the algorithm, it is independent of the selected ROI and more robust to INU in MRI images.

To the best of our knowledge, this is the first time a spatially variant image similarity term is introduced in a STAPLE framework, enabling the characterisation of both image similarity and human rater performance in a unified manner.

2. Methods

In this section, we first introduce the mathematical framework as presented in the original STAPLE algorithm by Warfield et al. (2004). We then introduce the idea of global and local ranking and the subsequent STAPLE model changes. Finally, we extend the full framework to a multi-label scenario.

2.1. The STAPLE algorithm

Let an image with N voxels be denoted by $\mathbf{y}$, with the intensity at voxel i denoted by y_i. Also, let $\mathbf{t}$ be an indicator vector of size N, again indexed by t_i, representing the hidden binary true segmentation of the object. The value of t_i will be equal to 1 when the structure is present in position i and equal to 0 when the structure is absent in position i. Let the $\mathbf{d}$ be a matrix of size $R \times N$, with each one of its rows $\mathbf{d}_j$representing a candidate segmentation of the object of interest obtained either by manual segmentation or an automatic algorithm. This row vector $\mathbf{d}_j$ has the same form as $\mathbf{t}$, with 1 and 0 representing the presence and absence of the structure at each position i. In order to parameterise the sensitivity and specificity of each rater, let $\mathbf{p} = (p_1, p_2, \ldots, p_R)^T$ and $\mathbf{q} = (q_1, q_2, \ldots, q_R)^T$ represent the sensitivity and specificity of each one of the R candidate segmentations, indexed by j. Here, $\mathbf{p}$ and $\mathbf{q}$ represent a global measure of agreement and disagreement, respectively, between a candidate segmentation and the consensus. Thus, they do not depend on the image index i. In order to estimate $\mathbf{t}$, one needs to maximise the log likelihood of the complete data of this problem $(\mathbf{d}, \mathbf{t})$ given the set of parameters $(\mathbf{p}, \mathbf{q})$. Thus, the cost function being optimised is the logarithm of the complete data likelihood $f(\mathbf{d}, \mathbf{t}|\mathbf{p}, \mathbf{q})$, described as

$$(\hat{\mathbf{p}}, \hat{\mathbf{q}}) = \arg\max_{\mathbf{p},\mathbf{q}} \log(f(\mathbf{d}, \mathbf{t}|\mathbf{p}, \mathbf{q})). \quad (1)$$

Using the definition of sensitivity and specificity, $\mathbf{p}$ and $\mathbf{q}$ can be described as the "true positive fraction" and "true negative fraction". Thus, p_j and q_j can be represented by

$$p_j = Pr(d_{ij} = 1 | t_i = 1),$$
$$q_j = Pr(d_{ij} = 0 | t_i = 0).$$

This model assumes that the candidate segmentations are independent from each other and thus $p_j \perp p_{j'}, q_j \perp q_{j'}$ and $d_{ij} \perp d_{ij'}, \ \forall j \neq j'$. Eq. 1 can thus be maximised by an Expectation–Maximisation algorithm. The notation w_i^k is used to represent the expected probability of the true segmentation at voxel i being equal to one at iteration k. Here, w_i^k is then defined as

$$w_i^k \equiv f(t_i = 1 | \mathbf{d}_i, \mathbf{p}^k, \mathbf{q}^k) \quad (2)$$
$$= \frac{a_i^k}{a_i^k + b_i^k} \quad (3)$$

with

$$a_i^k \equiv f(t_i = 1) \prod_j f\left(d_{ij} | t_i = 1, p_j^k, q_j^k\right) \quad (4)$$

$$b_i^k \equiv f(t_i = 0) \prod_j f\left(d_{ij} | t_i = 0, p_j^k, q_j^k\right) \quad (5)$$

and the parameters (p, q) at iteration $(k+1)$ are optimised by

$$p_j^{k+1} = \frac{\sum_i w_i^k d_{ij}}{\sum_i w_i^k} \quad (6)$$

$$q_j^{k+1} = \frac{\sum_i (1 - w_i^k)(1 - d_{ij})}{\sum_i (1 - w_i^k)}. \quad (7)$$

A more detailed explanation of the model derivation can be found in Warfield et al. (2004).

2.2. Iterative MRF regularization

Similarly to the original STAPLE algorithm, a Markov Random Field (MRF) is used to add spatial consistency. The MRF model presented in the original STAPLE paper is a post processing step that works on integer labels and not on the probabilities. In order to introduce the MRF spatial consistency within the same optimisation framework, the model presented in Cardoso et al. (2011a) is used. This model is not only computationally more efficient than the MRF model presented in the original STAPLE algorithm (Warfield et al., 2004), as it is updated with a mean field approximation, but it works on probabilistic labels and not on the final binarised labels.

This MRF model can be described as a non-binary multi-class extension of the Potts model with the neighbouring clique strength dependent of the voxel size. It has the form

$$f(t_i = k) = \frac{\pi e^{-\beta_i U_{\mathrm{MRF}(t_i = k)}}}{\sum_j \pi_j e^{-\beta_i U_{\mathrm{MRF}(t_i = j)}}}$$

with

$$U_{\mathrm{MRF}(t_i = k)} = \sum_{j=0}^{1} h_{kj} \left(\sum_{l \in \mathcal{N}_i^x} s_x w_{lj} + \sum_{l \in \mathcal{N}_i^y} s_y w_{lj} + \sum_{l \in \mathcal{N}_i^z} s_z w_{lj} \right)$$

where $\mathbf{H}$ is a $K \times K$ matrix with element h_{kj} containing the transition energy between the class k and the class j, $w_{lj} \equiv f(t_l = j | \mathbf{d}_l, \mathbf{p}^k, \mathbf{q}^k)$ and with the MRF neighbourhood system defined as $\mathcal{N}_i = \{\mathcal{N}_i^x, \mathcal{N}_i^y, \mathcal{N}_i^z\}$. Here, $\mathcal{N}_i^x$, $\mathcal{N}_i^y$, $\mathcal{N}_i^z$ represent the two direct neighbours of i in the x, y and z directions respectively. Also, π_k is the proportion of the object k in the full image, estimated from w at each iteration and s_x, s_y and s_z are the inverse of the voxel size in the x, y and z directions respectively. Note that π_k can be made spatially varying by using a Log-Odds framework as the one from Sabuncu et al. (2010). As the presented formulation only has two classes, the MRF matrix $\mathbf{H}$ is set up with the diagonal elements equal to 0 and the off-diagonal elements equal to 1.

When applying this MRF model in a multi-label fusion scenario, the MRF energy function can be extended to incorporate anatomically derived information about the expected neighbourhood transitions as in Cardoso et al. (2011a). Conversely, the classical MRF presented by Warfield et al. (2004) assumes that the transition between every pair of classes has the same probability. For the rest of the paper, β_i is considered constant throughout the image and equal to 0.5. Both the value of β_i and the matrix H can be optimised in order to improve the overall results. Nonetheless, we'll refrain from this optimisation due to it's computational complexity.

2.3. Global and region-of-interest based ranking

In the original STAPLE paper, Warfield et al. (2004) states that implicit in this model is the notion that the experts have been trained to interpret the images in a similar way. The segmentation decisions may differ due to random or systematic rater differences, and a probabilistic estimate of the true segmentation can be formulated as an optimal combination of the observed decisions and a prior model. Thus, these implicit assumptions may not hold when STAPLE is used for segmentation propagation. For segmentation propagation purposes, the errors can come from different morphological characteristics between the images, bad registration results and even the resampling method.

In order to ameliorate this problem, Aljabar et al. (2009) proposed the use of a global normalised cross correlation (GNCC) based metric to rank the registered templates according to the image being segmented in order to only include propagated segmentations that are consistently accurate. Leung et al. (2010) then introduced the same concept in a STAPLE framework, where the GNCC was calculated on a region of interest defined by the union of the propagated labels, resulting in an improved segmentation accuracy. This metric was used because it was shown to provide a good criterion for template selection in multi-centre imaging data (Aljabar et al., 2009). Once a rank of best to worst matches for each template was established, a subset of the highest ranked matchers was used to propagate the template labels onto the images to be segmented. This methodology still has some limitations, because the morphology of the structure and the quality of the registration is characterised as a single global image metric based on the NCC. Thus, in order to provide a good segmentation, either the registration algorithm must perform well in most cases, or the database has to have enough samples with the relevant type of morphology for the image being segmented. For example, if one wants to segment the temporal cortex of a patient's brain using segmentation propagation, the database would have to be large enough to contain enough templates with the same morphological features (e.g. number of sulci and giri) as the image to be segmented, so that the registration algorithm can match these features. Also, the registration might work very well in some areas but less well in other areas, leading to an ambiguous NCC value and to the introduction of errors in the label fusion process.

2.4. Local ranking for segmentation propagation

Without loss of generality, in this work, the local image similarity between images is assessed using the fast locally normalised correlation coefficient (LNCC), as proposed by Cachier et al. (2003). This choice of LNCC is contrary to what was suggested by Artaechevarria et al. (2009), as we have found better performance with a LNCC based image similarity than with the local mean

squared difference (LMSD). Nonetheless, the framework is general enough to allow any image similarity to be used.

The fast LNCC image similarity used in this work is similar to the LNCC presented by Artaechevarria et al. (2009) but the mean and standard deviation are calculated on a local Gaussian window using a convolution method. This makes the LNCC estimate smoother and computationally less expensive. Let $\mathbf{x}$ represent a propagated intensity image from the atlas after registration and $\mathbf{y}$ represent the target image to be segmented. Under this formulation, the LNCC at position i will be given by

$$LNCC_i = \frac{\langle \mathbf{y}, \mathbf{x} \rangle_i}{\sigma_i(\mathbf{y})\sigma_i(\mathbf{x})}$$

where

$$\langle \mathbf{y}, \mathbf{x} \rangle_i = \mu(\mathbf{y} \cdot \mathbf{x})_i - \mu(\mathbf{y})_i \cdot \mu(\mathbf{x})_i \qquad \mu(\mathbf{y} \cdot \mathbf{x})_i = \mathscr{G}_\sigma * (\mathbf{y} \cdot \mathbf{x})$$
$$\mu(\mathbf{y})_i = \mathscr{G}_\sigma * \mathbf{y} \qquad \mu(\mathbf{x})_i = \mathscr{G}_\sigma * \mathbf{x}$$
$$\sigma_i(\mathbf{y}) = \sqrt{\mu(\mathbf{y}^2)_i - \mu(\mathbf{y})_i^2} \qquad \sigma_i(\mathbf{x}) = \sqrt{\mu(\mathbf{x}^2)_i - \mu(\mathbf{x})_i^2}$$

with $\mathbf{y}^2$ representing the element-by-element squaring of $\mathbf{y}$, $*$ denoting the convolution operator, $\mathscr{G}_\sigma$ denotes a Gaussian smoothing kernel with standard deviation σ, and $\cdot$ denotes an element-by-element multiplication.

Due to the local nature and smoothness of the metric, the similarity between the images is described on a smooth voxel by voxel basis, enabling a voxel by voxel ranking with reduced discontinuity effect. If, for example, one starts from a set of 15 template images registered to the image under study, one can then calculate how much each one of the template images correlate locally with the image under study and then take only the top five templates on a voxel by voxel basis. There are three main advantages to the proposed method compared to using GNCC: first, the global nature of the GNCC metric may be an unrealistic assumption regarding the complex morphology and shape of the object under study. Secondly, it removes the need to create a ROI for GNCC calculation as the metric itself is local. Finally, this method is more robust to the presence of INU in MRI images, as the local nature of the LNCC method obviates the metric bias due to INU.

2.5. STAPLE with local ranking

In order to introduce this local ranking information in the previously described STAPLE algorithm, let a new model variable l_{ij} represent an observed cluster assignment that characterises the image similarity. For the sake of simplicity, l_{ij} will be equal to 1 if the image $\mathbf{g}_k$ is in the top X ranked images at position i and equal to 0 otherwise. Here, X controls the number of images to use locally according to the LNCC.

This new observation l_{ij} can be integrated into the STAPLE framework by altering the model to

$$(\hat{p}, \hat{q}) = \arg \max_{\mathbf{p},\mathbf{q}} \log(f(\mathbf{d}, \mathbf{t}, \mathbf{l}| \cdot \mathbf{p}, \mathbf{q})) \tag{8}$$

Using Jensen's inequality, the lower bound on the model will be given by

$$Q(\hat{p}, \hat{q}) = \sum_T \sum_i f(t_i|\mathbf{d}_i, \mathbf{l}_i, \mathbf{p}, \mathbf{q}) log[f(\mathbf{d}_i, \mathbf{l}_i|t_i, \mathbf{p}, \mathbf{q}) f(t_i)] \tag{9}$$

where $f(\mathbf{d}_i, \mathbf{l}_i|t_i, \mathbf{p}, \mathbf{q})$ is defined as a Bernoulli over a Bernoulli distribution

$$f(\mathbf{d}_i, \mathbf{l}_i|t_i, \mathbf{p}, \mathbf{q}) = \prod_j \left[f(d_{ij}|t_i, p_j, q_j)^{t_i} f(d_{ij}|t_i, p_j, q_j)^{(1-t_i)} \right]^{l_{ij}} 0^{(1-l_{ij})} \tag{10}$$

with t_i determining the true segmentation label and the observation l_{ij} determining if a template j is either a local morphological inlier or an outlier, i.e. if template j is similar or dissimilar to the target image after registration. As we are only interested in the subset of the data where $l_{ij} = 1$, a restricted maximum likelihood (REML) approach is used to focus on the likelihood of a subset of the data, where

$$w_i^k \equiv f(t_i = 1|\mathbf{d}_i, l_{ij} = 1, \mathbf{p}^{k-1}, \mathbf{q}^{k-1}) = \frac{a_i}{a_i + b_i} \tag{11}$$

with

$$\begin{aligned} a_i &\equiv f(t_i = 1) \prod_{j:l_{ij}=1} f(d_{ij}|t_i = 1, p_j^{k-1}, q_j^{k-1}) \\ &= f(t_i = 1) \prod_{j:\{l_{ij},d_{ij}\}=\{1,1\}} p_j^{k-1} \prod_{j:\{l_{ij},d_{ij}\}=\{1,0\}} \left(1 - p_j^{k-1}\right) \\ b_i &\equiv f(t_i = 0) \prod_{j:l_{ij}=1} f(d_{ij}|t_i = 0, p_j^{k-1}, q_j^{k-1}) \\ &= f(t_i = 0) \prod_{j:\{l_{ij},d_{ij}\}=\{1,1\}} q_j^{k-1} \prod_{j:\{l_{ij},d_{ij}\}=\{1,0\}} \left(1 - q_j^{k-1}\right). \end{aligned} \tag{12}$$

Here, p_j and q_j will now be

$$p_j^k = \frac{\sum_{i:l_{ij}=1} w_i^k d_{ij}}{\sum_{i:l_{ij}=1} w_i^k} \tag{13}$$

$$q_j^k = \frac{\sum_{i:l_{ij}=1} w_i^k (1 - d_{ij})}{\sum_{i:l_{ij}=1} w_i^k}. \tag{14}$$

In this REML framework, a_i, b_i, p_i and q_i are only influenced by the locations where $l_{ij} = 1$, i.e. only on the locations where the template image is locally similar to the image being segmented.

In this modification to the classic STAPLE algorithm, q_j and p_j now represent the sensitivity and specificity only in areas where $l_{ij} = 1$, i.e. each classifier is considered an expert by the LNCC ranking strategy. This results in a 2 step performance estimation that decouples the two sources of error: one based on the LNCC image similarity metric observation, modelled through l_{ij}, characterising the non uniform registration accuracy and shape differences, and the other step characterising the specificity and sensitivity of each classifier, through p_j and q_j, when compared with the consensus classification.

In this algorithm, we use a LNCC ranking-based binary cluster assignment for the observed variable l_{ij}. This approach is analogous to a sampling scheme, where samples with low local similarity are rejected from the fusion. However, the framework allows non-binary cluster assignments, where different samples can have different *importance* weights.

2.6. Performance parameter bias due to structure size

In the original STAPLE formulation, the performance parameters are estimated using all the samples from the image. More recent strategies by Rohlfing et al. (2004) and Asman and Landman (2011) have restricted the number of samples that are used to non-consensus areas in order to increase performance while reduce bias. Similarly, in this formulation and in STAPLE, if the size of the object and the size of the background are very different, the algorithms convergence results in both mathematical precision issues (due to the limited floating-point accuracy representation of $\mathbf{q}$ and $\mathbf{p}$) and biased performance parameters. For example, in a situation where the size of the object is much smaller than the background, the specificity q_j will tend to 1 because $\sum_i w_i(1 - d_{ij})$ will be approximately the same as $\sum_i w_i$ as most pixels in the image are $\mathbf{d}_i = 0$. Equally, due to the small size of the object, $\sum_i w_i d_{ij}$ will be much less similar to $\sum_i w_i$ and thus p_j will not be as close to 1 as q_j. This effect can be seen in Warfield et al. (2004) Table 1 and 2. When these biased values of p_j and q_j are then used to calculate the new w_{ij}, b_i will tend to 0, and thus w_{ij} will tend to 1. If the

STAPLE output w_{ij} is then thresholded at 0.5 confidence, the object will look over-segmented. In order to avoid an over-segmentation effect, one tends to threshold w_{ij} at very high values, *e.g.* a threshold of 0.9999, as used in (Leung et al., 2010). The optimal threshold will depend on many factors like the number of classifiers used, the mean value of **p** and **q** and even the value of β for the MRF. Also, because the value of *wik* will be very close but different from 1, numerical precision becomes an issue. Due to all these issues and given that this threshold is normally set to a constant value within the same study (even if more classifiers are used), the performance results of the STAPLE classifier fusion have a characteristic bumpy shape (Leung et al., 2010). Furthermore, the performance peak in terms of segmentation accuracy will depend on the chosen threshold, making all the analysis biased towards this choice.

Rohlfing et al., 2004 suggested that only updating and using disputed samples for parameter estimation can improve the computation time. One should note that this approach not only improves the computation time, but most importantly, it also improves segmentation performance by reducing the **p** and **q** unbalance and consequently the numerical precision issues. Thus, instead of trying to empirically set a threshold on the STAPLE probabilistic output, we restrict the parameter optimisation to non-consensus voxels. Thus, all the voxels where d_{ij} is equal to either 0 or 1 for all experts j are removed from the estimation. This method assumes that if all the classifiers agree on a label at a certain spatial position i, then the voxel is marked as solved and is not taken into account for the estimation of p_j and q_j. In this case, p_j and q_j represent the sensitivity and specificity only in ambiguous voxels, thus ameliorating the bias caused by structure size. One can then threshold w_{ij} at 0.5 without causing over-segmentation of the object. The effect of this step in terms of the shape of the performance results curve will be shown in the validation section.

In summary, the proposed method, named STEPS (Similarity and Truth Estimation for Propagated Segmentations), can be described as a combination of the LNCC ranking, the MRF and two STAPLE modifications regarding both the introduction of the local indicator function l_{ij} and the removal of consensus voxels from the parameter estimation.

2.7. Multi-label extention

Let **t** be an indicator vector of size N, indexed by t_i, representing the hidden true label describing several objects under analysis. This hidden label is denoted by an integer value $\{1, 2, \ldots, c\}$, with each value representing a different object of interest, from a total of c objects. Now, let **d** be a vector of size R, with each one of its elements $\mathbf{d}_r$ representing a candidate segmentation of the object of interest obtained either by manual segmentation or an automatic algorithm.

Table 2
Dice score statistics for hippocampal segmentation on 30 ADNI subjects using: STEPS, and the methods by Sabuncu et al., Yushkevich et al., Artaechevarria et al., all using the previously optimised parameters.

Fusion	STEPS	Sabuncu	Yushkevich	Artaechevarria	Spatial-STAPLE
Mean	0.903	0.870	0.875	0.874	0.880
SD	0.019	0.014	0.018	0.019	0.015
Median	0.907	0.870	0.877	0.875	0.879
p-Value	–	0.001	0.006	0.004	0.007

In order to extend the concept of sensitivity and specificity of a segmentor j into a multi-class model, a confusion matrix $\mathbf{N}_j$ and its row normalised equivalent λ_j, similar to the ones presented in Xu et al. (1992) and Rohlfing et al. (2004), are introduced in the notation. The matrix $\mathbf{N}_j$ is defined as

$$\mathbf{N}_j = \begin{pmatrix} n_{j11} & n_{j12} & \ldots & n_{j1c} \\ n_{j21} & n_{j22} & \ldots & n_{j2c} \\ \vdots & \vdots & \ddots & \vdots \\ n_{jc1} & n_{jc2} & \ldots & n_{jcc} \end{pmatrix}$$

with each element n_{jab} denoting that n samples of class a have been assigned a label b by segmentor j. The elements of the matrix λ_j are then defined as

$$\lambda_j(a,b) = \frac{n_{jab}}{\sum_c n_{jab}}$$

Similarly to Rohlfing et al. (2004), using the new definition of the performance parameter λ_j, the posterior probability for sample i to belong to class c will then be

$$w_{ia} = \frac{f(t_i = a)\prod_{j:l_{ij}=1}\lambda_j(a, d_{ij})}{\sum_c f(t_i = c)\prod_{j:l_{ij}=1}\lambda_j(c, d_{ij})}$$

and the performance parameter matrices λ_j are updated at each iteration by setting

$$\lambda_j(a,b) = \frac{\sum_{i:\{l_{ij},d_{ij}\}=\{1,b\}} w_{ia}}{\sum_{i:\{l_{ij}\}=\{1\}} w_{ia}}$$

In a multi-label scenario, instead of thresholding the output of w_{ia} at a certain value, the label with the highest value of w_{ia} at each position i is considered the optimal label.

Table 1
Leave-one-out cross validation statistics for different ranking methods and fusion approaches: STEPS, STEPS without MRF (STEPS-nMRF), STEPS with the MRF model proposed by Warfield et al. (STEPS-bMRF), STEPS including consessus areas (STEPS-cons), and the methods in Asman and Landman (2012), Aljabar et al. (2009), Yushkevich et al. (2010), Leung et al. (2010), Sabuncu et al. (2010) and Artaechevarria et al. (2009).

Fusion	STEPS	STEPS-nMRF	STEPS-bMRF	STEPS-Cons	Aljabar
Param.	$X = 15,\ \sigma = 1.5$	$X = 15,\ \sigma = 1.5$	$X = 15,\ \sigma = 1.5$	$X = 9,\ \sigma = 1.5$	$X = 17, D = 2$
Mean	0.925	0.919	0.920	0.921	0.907
SD	0.015	0.018	0.017	0.014	0.016
Median	0.929	0.918	0.919	0.922	0.909
p-Value	–	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
95% CI	–	0.004–0.007	0.004–0.009	0.004–0.006	0.016–0.020
Fusion	Leung	Asman	Yushkevich	Artaechevarria	Sabuncu
Param.	$X = 6, D = 2$	$w = 0.2,\ \kappa = 1$	$\alpha = 1, \sigma = 1.5$	$p = -5, r = 8$	$\sigma = 15, \rho = 1$
Mean	0.909	0.919	0.919	0.917	0.916
SD	0.015	0.015	0.013	0.014	0.013
Median	0.913	0.918	0.919	0.915	0.913
p-Value	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
95% CI	0.014–0.018	0.004–0.007	0.004–0.007	0.004–0.009	0.005–0.009

3. Validation

The validation of the proposed method is divided into two components, one for single label fusion and one for multi-label fusion. As the availability of manual segmentations is much greater for single labels, the validation of the proposed method for label fusion of individual structures was performed in five steps:

1. The method was applied to synthetic data to show the effect of STEPS on a simulated data set with different morphological properties.
2. We assessed the contribution of each improvement proposed in STEPS and then validated STEPS against other label fusion techniques using leave-one-out cross validation.
3. Using the optimised model parameters, a leave-one-group-out validation (jackknifing) was done to demonstrate robustness to simulated database size reduction.
4. Validation was then done on a subset of the publicly available ADNI database in order to show robustness to different atrophy states and imaging protocol.
5. STEPS was finally applied to the ADNI database to show volumetric group differences.

Due to the limited availability of template databases with multiple labels, only one validation step was performed for this scenario. Here, the performance of STEPS was compared to a set of state-of-the-art fusion techniques when segmenting a set of 30 brain images with 83 manually segmented structures using a leave-one-out cross validation.

3.1. Phantom validation

In order to validate the advantages of local ranking versus global ranking under a constrained experiment, a set of six simulated anatomical images with corresponding ground truth labels was generated. Each image represents a highly folded structure similar to the cortex, with the simulated intensities in line with anatomical T1 weighted MRI images. Rician noise was then added to the simulated anatomical images by adding Gaussian noise to both real and complex components in the Fourier domain. These six images have different number of gyri, representing different morphologies of the brain. One of these images was chosen as the image to segment and the other 5 were used as a template database. In order to simulate mis-registrations, 3 small random deformation fields were generated and applied per template (see Fig. 1(top right)), resulting in 15 different templates with 5 different morphologies, each one with a corresponding label. The proposed method's segmentation was compared to the GNCC-ranked STAPLE (Rohlfing et al., 2004,Yushkevich et al., 2010) using the Dice score as a performance metric. For both the method proposed by Leung et al. and STEPS, we took the top five templates ranked globally (according to the GNCC) and locally (according to the LNCC) respectively.

Results are shown in Fig. 1. Using a leave-on-out cross validation, the mean Dice score for STEPS and the Leung et al. based method was 0.939 and 0.753 respectively.

3.2. Hippocampal segmentation

This section validates the performance of the proposed technique for hippocampal segmentation. A previously described hippocampal template library of manually segmented regions, from 55 subjects, was used (Barnes et al., 2008). The subjects in the template library included 36 subjects with clinically diagnosed AD and 19 controls who had a mean age of approximately 70 years. All scans were acquired at a single site 1.5 T GE scanner using a volumetric T1-weighted sequence. The left and right hippocampal regions were segmented by an expert segmentor. In order to increase the template database size, each image and its flipped mirror image were used as templates, resulting in 110 templates with associated segmentations.

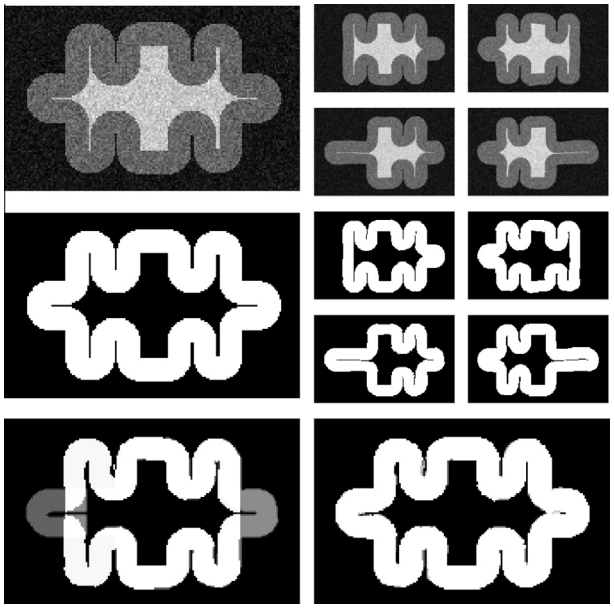

Fig. 1. From left to right: (Top) The image to segment, four samples from the simulated template database. (Centre) The ground truth segmentation and the respective labels from the template database with different morphologies and simulated registration errors. (Bottom) The probabilistic segmentation using the method from Leung et al. (left) and STEPS (right). Note that the lack of local matching has limited the ability of the GNCC method to capture the local features due to the morphologically restricted database.

In order to assess STEPS, we performed a leave-one-out segmentation validation on all the images. For each image, the remaining 108 templates from the other 54 subjects were used in order to minimise bias due to same-subject left–right hippocampal symmetry. Each template was first affinely registered (12 DOFs) using a block matching approach (Ourselin et al., 2000; Ourselin et al., 2001) and then non-rigidly aligned using a fast free-form registration algorithm (Modat et al., 2010) to the image under study. The resulting transformations were used to propagate the manual segmentations to the image under study and resampled using nearest-neighbour interpolation in order to maintain their binary nature.

This section of the validation has two main purposes. First, is to validate the influence of each component of the method, i.e. LNCC, MRF and consensus voxel removals. In order to do so, the proposed method was compared with the ROI normalised cross correlation (ROINCC) based ranking under a majority voting and STAPLE fusion strategies as proposed by Aljabar et al. (2009) and Leung et al. (2010) respectively, thus assessing the merit of adding the local ranking strategy. The proposed MRF model was also compared to the model proposed in (Warfield et al., 2004) (here referred as STEPS-bMRF), in order to test the merit of the iterative and probabilistic MRF. Finally, the influence of removing consensus areas was tested by running STEPS including consensus voxels.

The second purpose of this validation section is to validate the performance of STEPS against state-of-the art methodologies. Thus, STEPS was also compared with spatial-STAPLE (Asman and Landman, 2012), the method by Sabuncu et al. (2010), LNCC weighted voting presented in Yushkevich et al. (2010) and the MSD weighted voting presented in Artaechevarria et al. (2009). All methods were implemented as part of the NiftySeg package, except the Spatial-STAPLE and the method by Sabuncu et al. (2010), where we use

the implementations provided by the authors, available at http://masi.vuse.vanderbilt.edu and http://people.csail.mit.edu/msabuncu/ respectively.

Note that all these comparisons only test the merit of the fusion strategy and not the performance of the full pipeline, as all the templates are registered in the same manner. the code provided by Sabuncu et al. (2010) was modified, as suggested by the author, in order to accept the same registration strategy.

3.3. Parameter optimisation and algorithm comparison

In order to optimise the fusion parameters, the Dice score between the estimated segmentation and the manual segmentation was calculated for different values of Gaussian kernel size, number of labels used and registration parameters. The parameters for all the other methods were also optimised.

For all other methods based on ranking, we took the top X ranked images, with X varying between 3 and 25. Only odd numbers of X were used in majority voting to avoid bias due to voting ties. For the LNCC ranking the images were locally ranked by setting $l_{ij} = 1$ if the registered template k was in the top X ranked images at position i and to 0 otherwise. For the LNCC ranking in STEPS the value of σ was varied between 1 mm and 2 mm with an increment of 0.25 mm and between 2 mm and 6 mm with an increment of 1 mm, for each value of X, in order to find the optimal Gaussian kernel size. Regarding the other parameters, the region of interest in Aljabar et al. (2009) and Leung et al. (2010) was defined as the union of all the propagated labels dilated D times. The parameter D was also optimised. For each value of X, D was varied between 1 and 4.

The parameters for the methods that are not based on ranking, like the method proposed by Asman and Landman (2012); Yushkevich et al. (2010); Artaechevarria et al. (2009); Sabuncu et al. (2010), were also optimised. As suggested in the original paper, the Spatial-STAPLE window size w was varied between 0.1 and 0.3 (sampling spacing 0.1), the global performance level bias κ was assessed at samples 0.1, 1 and 10 and the overlap between windows was set to 0.5. For Yushkevich et al. (2010), the value of α was varied between 0.5 and 2 (sampling rate 0.5) and σ was varied between 0.5 and 2.0 (sampling rate 0.5). For Artaechevarria et al. (2009), the value of p was varied between -10 and 10 (sampling rate 5) and r was varied between 4 and 16 (sampling rate 4). Finally, the parameters for Sabuncu et al. (2010) were optimised, also as suggested in the original paper, with σ varying between 5 and 15 (sampling spacing 5) and ρ varying between 0.5 and 1.5 (sampling spacing 0.5).

The registration parameters were not optimised within the same scheme due to computational complexity. They were only visually optimised on a subset of 10 images in order to produce good registration accuracy. The optimal registration parameters were found to be 2.5 mm control-point spacing with 1% bending energy as regularisation. Due to the overestimation explained in Section 2.6, a constant threshold of 0.9999 was used for all the STAPLE based methods, in order to obtain the final binary segmentation. This threshold is identical to the one used in Leung et al. (2010). For all other methods, the threshold was set to 0.5 due to their unbiased nature.

In order to assess the accuracy of the segmentation, the Dice score was calculated between the ground truth manual segmentation and the obtained binary segmentation. An example segmentation from STEPS is shown in Fig. 2 and the Dice scores for different parameters using STEPS are shown in Fig. 3. The optimal parameters are shown in Fig. 4. For STEPS, STEPS without the proposed MRF regularisation and STEPS with the MRF model proposed in

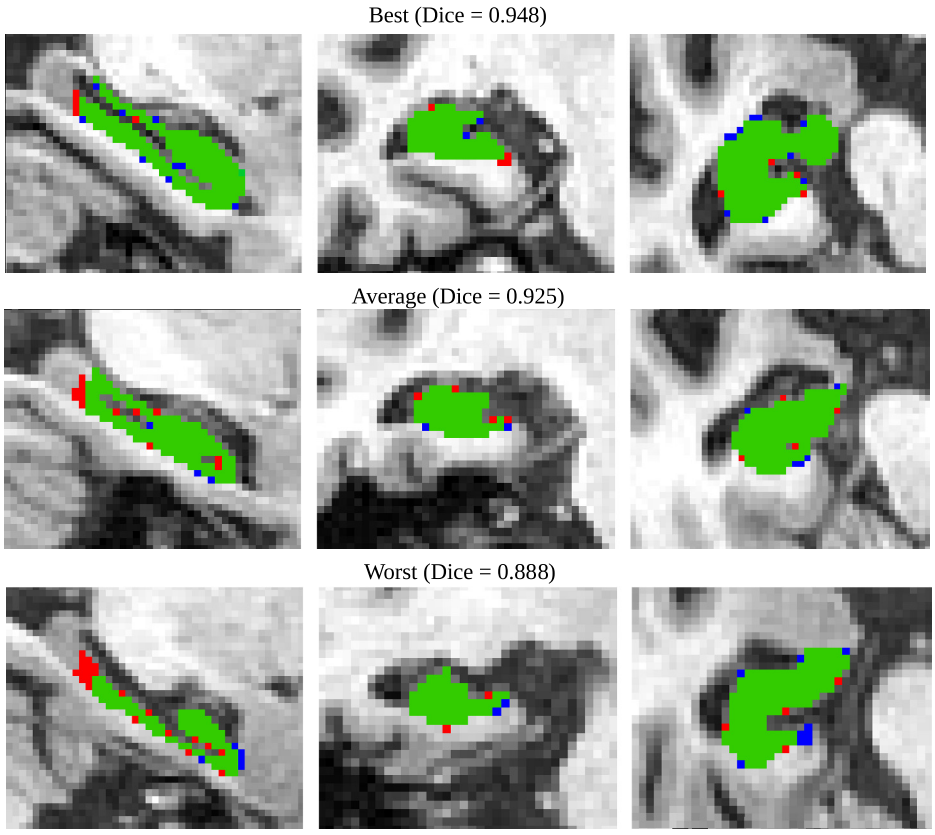

Fig. 2. Segmentation results showing the best, an average and the worst result. The blue, red and green colours represent the ground truth, the proposed method and the overlap between both segmentations respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

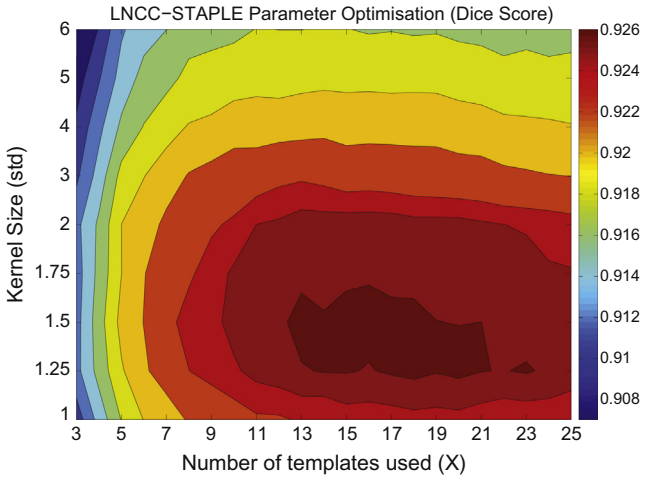

Fig. 3. Mean Dice score for varying values of σ and X for the proposed STEPS method using a leave-one-out cross validation. The best parameters were found to be $X = 15$ and $\sigma = 1.5$, with a mean Dice score of 0.925 for STEPS

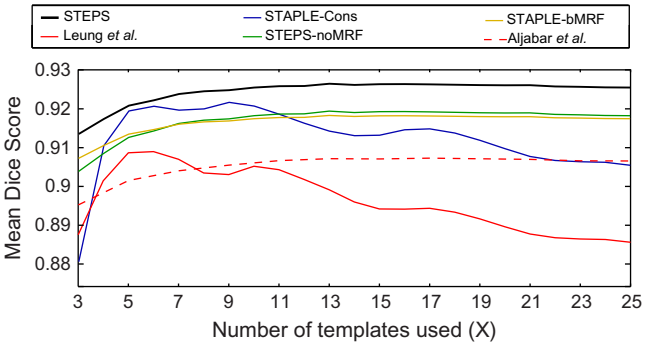

Fig. 4. The mean Dice score for all methods based on ranking, when applied to the full data set for varying values of X with optimal σ and D parameter. Labels are described in Table 1.

(Warfield et al., 2004), the optimal parameters are $X = 15$ and $\sigma = 1.5$ (Mean Dice = 0.925, 0.919 and 0.918 respectively). The parameters $X = 9$ and $\sigma = 1.5$ (Mean Dice = 0.921) are optimal for STEPS without the consensus voxels rejection as in for the fusion approach, as in Cardoso et al. (2011). The parameters $X = 6$ and $D = 2$ (Mean Dice = 0.909) and $X = 17$ and $D = 2$ (Mean Dice = 0.909) are optimal for the fusion approach in Leung et al. (2010) and in Aljabar et al. (2009). The best parameters for the Spatial-STAPLE method proposed byAsman and Landman, 2012 was $\{w, \kappa\} = \{0.2, 1\}$ (Mean Dice = 0.914). Finally, the parameters for the fusion approaches in Yushkevich et al. (2010); Artaechevarria et al. (2009) and Sabuncu et al. (2010) were found to be $\{\alpha, \sigma\} = \{1, 1.5\}, \{p, r\} = \{-5, 8\}$ and $\{\sigma, \rho\} = \{15, 1\}$ respectively. The mean Dice score was 0.919, 0.917 and 0.916 respectively. These optimal parameters are used for all comparisons. The Dice score statistics for all methods are shown on Table 1 and Fig. 4.

Using a two tail unequal variance paired t-test, STEPS performed significantly better ($p < 10^{-4}$) than all the other ranking and label fusion strategies for hippocampal segmentation. Confidence intervals for the mean differences, shown in Table 1, were found assuming normality of the paired differences. Interestingly, the standard deviation of the Dice score did not increase between STEPS and the regionally ranked fusion algorithms. STEPS achieves very high Dice score (0.907) for the 10th percentile data, with the worst segmentation having a Dice score of 0.888. For comparison, the method by Leung et al. (2010) and by Aljabar et al. (2009) only achieved Dice scores of 0.886 and 0.890 respectively for the 10th percentile and a Dice scores of 0.819 and 0.830 respectively for the worst segmentation. Furthermore, the proposed method (STEPS) has a Dice score equal or higher than all other methods for all data sets. The methods by Yushkevich et al. (2010); Artaechevarria et al., 2009; Sabuncu et al., 2010 and Spatial-STAPLE all show improved results when compared to both Leung et al. (2010) and Aljabar et al. (2009). However, when compared to STEPS, they still perform significantly worse.

3.4. Robustness to database size reduction

One of the main caveats of global ranking methods is the implicit necessity to have a large database in order to be able to represent the population's global anatomical variability. Conversely, STEPS describes image similarity on a local manner. Intuitively, this means that fewer templates are needed to describe the global anatomical variability of a population, as each template contributes locally to the global anatomical variability.

In order to test this hypothesis, we used the same data set as before. However, instead of using a leave-one-out approach, we used a subset of the available template database (110 templates) by selecting a smaller set of templates randomly (jackknifing). This is done in order to study the effect of reducing the size of the template database on the results. Assuming a simulated template database of size R, for each data set in the original database, 10 sets of R samples were randomly selected from the remaining 108 templates from 54 subjects. Each one of these 10 sets was then considered as a simulated database of size R used to segment the data set under study. The optimised parameters described in Section 3.2 were used in order to obtain the fused segmentations. For the sake of comparison, STEPS was compared to the method by Leung et al. (2010) and also to STEPS without excluding the consensus areas (STEPS-Cons).

The degradation was tested at three different levels of R (30, 60 and 90), with X varying between 5 and 25 (sampled only at odd values). Thus, 36,300 fusions were performed for each method, producing 10 segmentations per data set, per value of X and per value of R. The resulting Dice score are presented in Fig. 5. Using an unequal variance paired t-test to compare the Dice scores, STEPS performed significantly better ($p < 10^{-4}$) using only 30 templates than the ROINCC method using the full database.

3.5. Validation on a subset of the ADNI database

In order to characterise the accuracy of using a predefined template database to segment data sets from another database, an expert segmentor manually delineated the left hippocampus on the baseline and repeat T1-weighted MR images of 30 randomly selected subjects (IDs available in Appendix B). The data consists of 10 Alzheimer's disease (AD), 10 Mild Cognitive Impairment (MCI) and 10 controls, from the ADNI data set. Representative imaging parameters were TR = 2400 ms, TI = 1000 ms, $TE = 3.5$ ms, flip $angle = 8^{\circ}$ with either a $1.25 \times 1.25 \times 1.2$ mm^3 or a $0.94 \times 0.94 \times 1.2$ mm^3 voxel resolution. The T1-weighted volumetric scans were already pre-processed using the standard ADNI pipeline, including post-acquisition correction of gradient warping, B1 and INU correction and phantom based scaling correction.

The optimised parameters obtained above were used to segment this subset of the ADNI dataset. Segmentation accuracy was accessed by calculating the Dice score between the manual and automated segmentations. Results are shown in Table 2. Statistical differences were calculated using a two tail unequal variance paired t-test. The mean (SD) Dice score for STEPS was 0.903 (0.021), significantly higher than all the other methods (Yushkevich et al., 2010; Sabuncu et al., 2010; Artaechevarria et al., 2009

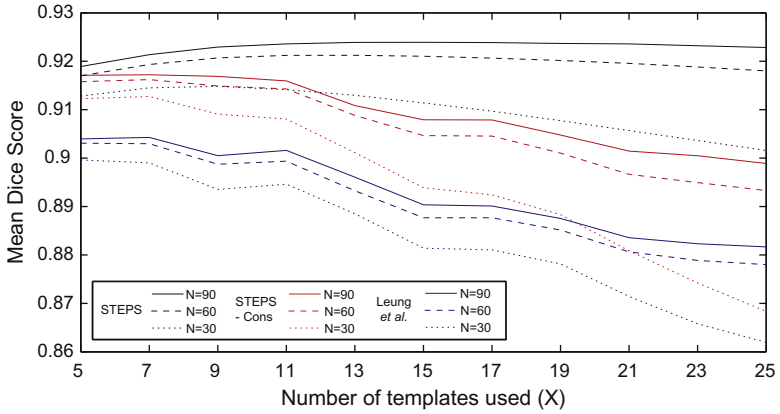

Fig. 5. The mean Dice score for varying values of X on a simulated database of size $R = 90$, $R = 60$ and $R = 30$ for STEPS, the method by Leung et al. (2010) and STEPS without excluding the consensus areas (STEPS-Cons). Note that there is almost no performance deterioration for the STEPS algorithm between a database of size 90 and 60. STEPS also performs significantly better using only 30 templates than the method proposed by Leung et al. (2010) using the full database.

and Spatial-STAPLE) at $p < 0.01$. The means, std, medians of the Dice scores and the p-values when compared to STEPS results are available in Table 2.

3.6. Hippocampal measures on the full ADNI data-set

In this section, the ADNI data sets were used to assess both hippocampal volume and change in volume over time (atrophy rate). As suggested in Lötjönen et al. (2011), in order to add PV information to the binary hippocampal segmentation and thus increasing statistical power, each image was also segmented using LoAd (Cardoso et al., 2011a). Hippocampal volume was considered as the sum of the GM fractional content at each voxel position within the binary segmentation obtained from STEPS, multiplied by the voxel size. The volumes of the left and right hippocampi were added together to give "total" hippocampal volume for each subject.

For the sake of comparison with previously published studies, only the baseline and 12-month repeat volumetric T1-weighted MR scans acquired using 1.5 T scanners were used. In total, 682 subjects were used (200 controls, 335 MCI and 147 AD). The scans were pre-processed following the standard ADNI pipeline, summarised in Leung et al. (2010). Demographics are shown in Table 3.

Linear regression was used to assess differences in volumes and change in volumes across groups. The volume, calculated as described above, is considered as dependent observed data. For cross-sectional analysis, the metadata available from the ADNI database comprising of age and gender was used as independent confounding variables. The total intracranial volume (TIV), obtained automatically using SPM8 as described in Leung et al. (2010), was also considered as a confounding variable. For the longitudinal assessment, the atrophy rate was estimated by measuring the difference in volume between baseline and repeat scans normalised by the baseline scan. Because the number of days between baseline and 1-year scans was different between subjects, this information was additionally used as a confounding variable.

Table 3
Subject demographics of the ADNI data set. Mean (SD) unless specified otherwise.

	Controls	MCI	AD
# data sets	200	335	147
Gender, # male	106	213	78
Age, years	76.0 (5.1)	74.9 (7.2)	75.3 (7.3)
Scan. Interval, days	396.3 (46.0)	396.3 (24.3)	390.1 (22.6)
TIV, ml	1584 (144)	1567 (149)	1554 (154)

The results are shown in Fig. 6. Statistical differences were calculated using a two tail unequal variance t-test and the significance level was set to $p < 10^{-3}$ due to the intrinsic pathological variability.

The cross-sectional study shows statistically significant hippocampal volumetric differences between the different disease groups. The mean volumes were also similar to previously estimated manual and automatic volumes. For the longitudinal study, even though atrophy rates were not derived directly from the registered serial MR images or propagated from baseline to repeat, the accuracy of the proposed method enables a direct comparison between the volumes of the hippocampus at baseline and 12-month follow up. Results shown in Fig. 6 and Table 4 show statistically significant differences in the mean atrophy rate between disease groups.

3.7. Multi-label segmentation propagation and comparison

The limited availability of template databases with multi-label s does not allow as complex a validation as with the single label scenario. Thus, only one leave-one-out cross validation was performed, making the validation anecdotal for untested morphologies and severe pathological cases. A previously described template library of 83 manually segmented regions from 30 subjects was used (Hammers et al., 2003; Hammers et al., 2007). The median age of all subjects was 31 years, ranging from 20 to 54 years, equal gender distributions and 83% right handed subjects. Scanner parameters are described in Hammers et al. (2007). In order to assess the accuracy for brain using STEPS, we performed a leave-one-out segmentation validation on all the datasets. Each image was first skull stripped using the method proposed by Segonne et al. (2004). Then, for each one of the 30 datasets, the remaining 29 templates were first affinely registered (12 DOFs) using a block matching approach (Ourselin et al., 2000) and then non-rigidly aligned using a fast free-form registration algorithm (Modat et al., 2010) to the image under study. The manual segmentations were then propagated using the previously estimated transformations and resampled using nearest-neighbour interpolation in order to maintain their binary nature. We compare STEPS to a previously published state-of-the-art method called MAPER (Heckemann et al., 2010); Yushkevich et al., 2010; Sabuncu et al., 2010 and Artaechevarria et al., 2009. STEPS was also tested without the MRF in order to show the improvements in accuracy and smoothness. In order to provide a fair comparison between methodologies, all the methods were compared using the same registration strategy. As the results from MAPER are highly dependent on

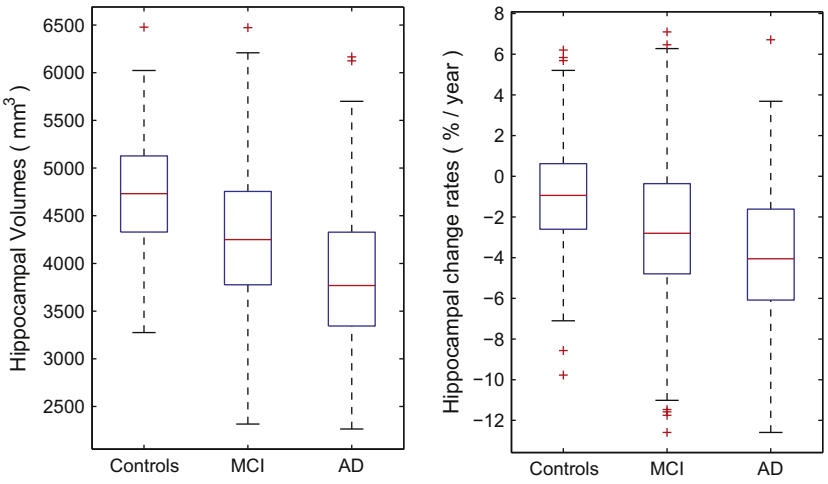

Fig. 6. Cross-sectional and longitudinal study on 682 data sets from the ADNI database. Left: Total hippocampal volume (left + right side) at baseline; Right: Hippocampal atrophy per year as a percentage of the baseline volume.

Table 4
Hippocampal volumes and change rates.

	Controls	MCI	AD
Volumes (mm^3)			
Mean	5195	4786	4427
Median	5152	4733	4218
SD	656	781	903
Change rates (%/year)			
Mean	1.09	2.74	4.04
Median	0.98	2.61	3.95
SD	3.0	3.5	3.6

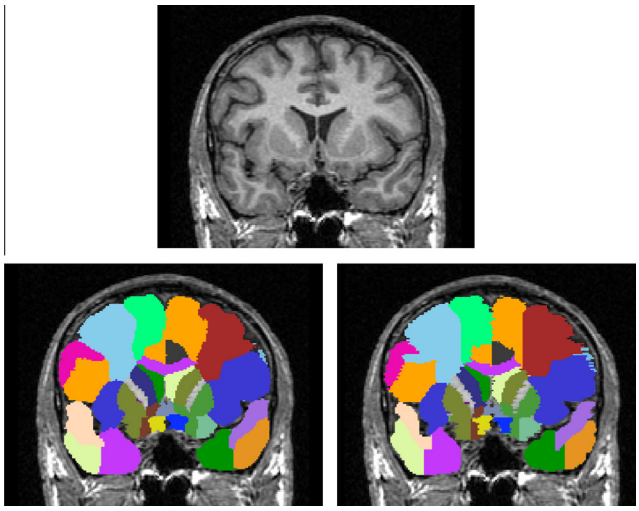

Fig. 7. An example showing the template (top), the automated Multi-STEPS segmentation (bottom-left) and the manual segmentation (bottom-right). Note the smoothness of the boundaries for the automated segmentation method.

the registration strategy, the results presented for the MAPER algorithm were kindly provided to us by the author. Results are shown in Table 5 and Fig. 7.

Results show that STEPS with the MRF outperforms the other techniques in many key internal structures. More specifically, STEPS outperforms the methods by Heckemann et al. (2010) (12 out of 83 structures), Yushkevich et al. (2010) (17 out of 83 structures), Sabuncu et al. (2010) (20 out of 83 structures) and Artaechevarria et al. (2009) (18 out of 83 structures) at $p < 10^{-4}$. The putamen was the only structure where another fusion algorithm (MAPER) outperformed STEPS, but that difference was not statistically significant ($p = 0.02$).

The MRF introduced in this model not only results in a segmentation accuracy improvement but also improves the smoothness of the boundary between the labels. Anatomically, each one of the parcelated areas should have one single connected component. In order to test the advantages of introducing the MRF into the algorithm with regards to discontinuity, the average number of connected components per parcelated area and per subject was calculated for the proposed method with and without the MRF and for the manual. The average (SD) number of connected

Table 5
Mean Dice coefficient for each structure, comparing the proposed method with and without MRF and MAPER (Heckemann et al., 2010). For bilateral structures, the left and right Dice score is averaged. Results for a set of key internal grey matter structures are shown.

Structure name	Fusion method					
	STEPS	STEPS no MRF	MAPER	Yushkevich	Sabuncu	Artaechevarria
Hippocampus	0.842	0.840*	0.828**	0.832*	0.820**	0.835*
Amygdala	0.805	0.803	0.789**	0.788**	0.775**	0.795*
Caudate Nucleus	0.892	0.890	0.891	0.887**	0.887**	0.877**
Nuc. Accumbens	0.695	0.687*	0.682**	0.680**	0.667**	0.690*
Putamen	0.891	0.888	0.894	0.890	0.874**	0.890
Thalamus	0.894	0.892	0.887**	0.883**	0.886**	0.878**
Globus pallidus	0.798	0.793*	0.771**	0.770**	0.773**	0.773**

* Significantly higher Dice scores are shown, with ** representing $p < 10^{-3}$.
** Significantly higher Dice scores are shown, with representing $p < 10^{-4}$.

components per parcelated area was 8.9(1.3), 13.7(1.8), and 4.08(0.4) for STEPS with and without the MRF and for the manual respectively. A Welch's two-tailed paired t-test was performed in order to test significance. These results show a statistically significant ($p < 10^{-4}$) reduction in the mean number of connected components per parcelated area when comparing STEPS with the MRF to STEPS without the MRF. STEPS with the MRF still performs significantly worse than the manual segmentation with regards to discontinuity and fragmentation of parcelated areas.

4. Discussion

In this work, we have developed an extension of the popular STAPLE algorithm that uses local intensity features to select the best labels to fuse, a novel iterative MRF to ensure spatial consistency and an uncertainty ROI optimisation to un-bias the algorithm towards larger structures. Both the robustness and accuracy of the segmentation were evaluated on the training set and in an independent database of cross-sectional and longitudinal brain MRI scans and tested the ability to directly use the segmentation for volumetric and atrophy rate measurements.

The algorithm was first tested on a simulated phantom with known ground truth segmentation, as a proof of concept. The STEPS method performed better than the STAPLE-GNCC method, presumably due to uncertainty caused by the lack of images in the template database with the same overall morphology as the image being segmented. Conversely, STEPS achieves a good overall segmentation due to the finer anatomical scale of the metric, suggesting that STEPS may enable the use of smaller template databases to describe the full population variability, leading to an improvement in both accuracy and computation time.

The proposed method was then applied to clinical data for the purpose of segmenting hippocampi. In order to find the parameters that produce the most accurate segmentations, a leave-one-out cross validation strategy was used to sample the overall accuracy from the parameter space. Each component of the proposed method was tested independently in order to assess the contribution of each one of the changes. To this end, we tested the contribution of the local ranking against global and regional ranking, the contribution of the proposed MRF model against the model proposed by Warfield et al. (2004) and STEPS without MRF, and also the contribution of the uncertainty ROI selection. The proposed method was then compared to publicly available techniques: STAPLE, Spatial-STAPLE and the methods by Yushkevich et al. (2010); Sabuncu et al. (2010) and Artaechevarria et al. (2009). Visual (see Fig. 2) and quantitative assessment demonstrates good segmentation accuracy and robustness, with the worst segmented image having a Dice score of 0.888. The improvements proposed in STEPS all provide significantly advantages ($p < 10^{-4}$) demonstrating the advantage of combining the local ranking, the new MRF model, the ROI optimisation strategies and the rater performance model. It also performs significantly better than all the other tested fusion techniques. Furthermore, the proposed method obtains a (mean $\pm$ *SD*) Dice score (0.925 ± 0.021) close to the inter-rater variability of the manual segmentors (0.93 ± 0.03), assessed on a different database (Leung et al., 2010).

Another advantage of local ranking strategies is that they implicitly encode local morphological variability rather than global morphological variability. Fewer anatomical templates are needed to deal with the population's overall morphological variability. In order to test this idea, a second experiment was performed in order to show that local ranking can still obtain the higher segmentation accuracy as global ranking when using fewer anatomical templates. This is advantageous because if one can represent complex shapes with fewer samples, the need for a large and accurate template database is greatly reduced. The results of the jackknifing shown in Section 3.4 demonstrate that STEPS can obtain significantly better segmentation accuracy, when measured using the Dice score, than the STAPLE-ROINCC label fusion algorithm, even when using three times fewer templates. As expected, there is a small shift of the Dice score peak for the optimal value of X between different database sizes with the optimal value of X shifting to higher values with an increase in database size. Another interesting fact, not present in Fig. 5 is the consistent and significant reduction, when compared to STEPS - Cons and Leung et al., in the standard deviation of the Dice score per data set after the 10 simulations. This means that the proposed STEPS method not only produces better results but is also less dependent on the choice of data sets the template database is composed. This is important in situations where no knowledge is available about the morphology of a population or when the database size in inherently small.

This extra robustness with regards to database size can be exploited to improve computational efficiency. One can enforce morphological sparseness of the template database by learning the manifold structure of the data from a set of deformation fields to a group-wise space. This sparse representation of the morphological characteristics of the population would greatly reduce the computational complexity without degrading the segmentation accuracy. One should note that this effect was validated only on AD, MCI and controls using a template library based of AD and controls. It remains to be seen if results hold for hippocampi with different atrophy patterns and different intensity profiles such as in hippocampal sclerosis and certain atrophy syndromes like frontotemporal lobar degeneration and semantic dementia.

All experiments summarised above were performed on the training set using either leave-one-out cross validation or jackknifing. To test the performance of the fusing strategy on data from a different database acquired with a different MRI imaging systems and protocols, the same label fusion techniques were also used to segment a subset of data from the ADNI database with manual segmentations. Using the parameters optimised in Section 3.3, STEPS achieved a Dice accuracy above 0.9, significantly higher ($p < 10^{-3}$) than all the other fusion methods, mainly due to the limited sample size. One can argue that the flat and larger plateau in the parameter selection of STEPS makes the segmentation less sensitive to changes in the imaging protocol, contrast and noise.

In a single label scenario, the STEPS algorithm was finally used to segment the hippocampi of all 682 1.5T ADNI data sets at baseline with 12-month repeat. Using the baseline data for a cross-sectional study, the volumetric results described in Section 3.6 show the expected significant separability in terms of volume, between AD, MCI and controls. Using both the baseline and 12-month repeat in a longitudinal study, the results show again significant group discrimination between AD, MCI and controls. The atrophy rates are in line with those previously reported, with a mean hippocampal atrophy rate (%/year) of 4.04, 2.74 and 1.09 for the AD, MCI and control subjects respectively. These results were achieved using volumetric data from the binary hippocampal segmentations combined with tissue segmentation. We hypothesise that should baseline and followup scans be treated non-independently with regards to the template propagation or if the measurement of atrophy was changed to the boundary shift integral (Leung et al., 2010), our longitudinal measures would reduce in terms of noise or variability with possibly improved disease group separation.

Lastly, in a multi-label propagation scenario, the algorithm was tested against the same set of fusion techniques. Results showed significant increases in segmentation performance, mainly in key internal grey matter structures like the hippocampus, amygdala, thalamus, globus pallidus and nucleus accumbens, known to be associated with several diseases. Furthermore, the statistically significant reduction in the number of connected components per

structure shows the advantage of using STEPS with the MRF spatial smoothness term when compared to STEPS without MRF. Due to the locality of the similarity metric, we also speculate that the proposed methodology should provide improvements in the of pathological subjects and patients with different brain morphologies. However, further validation of multi-atlas based brain is necessary as the current findings are anecdotal for untested morphologies and pathological cases. This is specifically important in pathological situations that lead to large anatomical deformations (*e.g.* ventriculomegaly, highly atrophied brains), as some of these morphological changes might no be correctly captured by the non-rigid image registration step. In order to reduce the complexity and consequently the errors of the mapping between morphologically dissimilar images, the segmentations can be propagated via morphologically similar intermediate datasets using an approach similar to the one proposed by Wolz et al. (2010) and Cardoso et al. (2012).

The current limitations of the proposed work are mostly related with the similarity metric. As previously described, even though the LNCC metric has many advantages when compared to a global metric, the local support of the metric can be problematic in low contrast areas. For example, if the non-rigid mapping between a normal subject and an AD patient with enlarged ventricles does not perform well enough, an area in the patient's ventricular cerebrospinal fluid can be mapped and will correlate very well with the white matter area in the normal subject. This problem is caused by the local normalisation of the mean intensity between the two regions and can be ameliorated by a multi-level version of the same metric or by combining both local and global similarity metrics. Furthermore, as suggested by Souvenir and Pless (2007) and Cardoso et al. (2012), the local intensity similarity metric can also be augmented by a morphological similarity metric based on the local displacement between mapped regions, thus introducing knowledge about anatomical shape changes. Nonetheless, the proposed framework is general enough and allows the replacement of the LNCC metric by any other similarity metric. Finally, because the proposed method and Spatial-STAPLE share the same construction and because they both seem to model rater and registration errors quite well, we believe that mixing both the proposed strategy and the spatially varying rater performance estimation of Spatial-STAPLE, i.e. calculating spatially variant performance parameters only in areas with uncertainty, would provide further improvements in the accuracy of the method.

In this paper, the focus has been on improving both the accuracy and robustness of segmentation propagation techniques by improving the label fusion component. Nonetheless, the algorithm's accuracy is still dependent, though to a smaller degree, on the quality of the manual segmentations and the type of pathologies and atrophy patterns represented in the template database. Further validation is still necessary in order to enable the unsupervised use of this algorithm in a clinical setting and for different disease groups. Additionally, the manual segmentation protocols can also be improved in order to avoid arbitrary cutoffs of structures, like the tail of the hippocampus, which may negatively affect the algorithm accuracy.

5. Conclusion

This paper presents a new algorithm, called STEPS, that incorporates a fast locally normalised cross correlation (LNCC) based ranking combined with a consensus based ROI selection and a new iterative MRF into the STAPLE formulation. The algorithm was first tested on a database of manually segmented hippocampi using a leave-one-out cross validation. Results show a significant improvement in terms of Dice overlap when compared to state-of-the-art label fusion algorithms, achieving a mean Dice score of 0.925. The STEPS label fusion technique also achieved better accuracy than globally ranked techniques even when using only a third of the templates, diminishing the necessity of large template databases. When tested on an independent database with data sets from different MRI imaging systems and protocols, STEPS still achieved an average Dice score above 0.9, again significantly higher than other techniques. Furthermore, cross-sectional and longitudinal hippocampal volumetric studies showed expected significant differences in volume and atrophy rates between AD, MCI and controls. Finally, when applied to multi-atlas segmentation propagation, STEPS showed a statistically significant increase in segmentation accuracy in several key brain structures when compared to MAPER and the methods by Yushkevich et al. (2010); Sabuncu et al. (2010) and Artaechevarria et al. (2009).

Acknowledgments

This work was undertaken at UCL/UCLH which received a proportion of funding from the Department of Health's NIHR Biomedical Research Centres funding scheme. The Dementia Research Centre is an Alzheimer's Research Trust Co-ordinating centre and has also received equipment founded by the Alzheimers Research Trust. M.J.C. is supported by a scholarship from the Fundação para a Ciência e a Tecnologia, Portugal (Scolarship number SFRH/BD/43894/2008) and by the EPSRC Program grant (EP/H046410/1). KKL acknowledges support from the MRC, ARUK and the NIHR. J.B. is supported by an Alzheimer's Research Trust (ART, UK) Research Fellowship partly supported by the Kirby Laing Foundation. N.C.F. is funded by the Medical Research Council (UK). S.O. is funded by both the EPSRC (EP/H046410/1) and the CBRC Strategic Investment Award (Ref. 168).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; 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.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; 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 organisation 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 California, Los Angeles. This research was also supported by NIH Grants P30 AG010129, K01 AG030514, and the Dana Foundation. The authors thank the ADNI study subjects and investigators for their participation.

Appendix A. Clinical data

Data used in the preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). The ADNI was launched in 2003 by the

National Institute on Ageing (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organisations, 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 Alzheimers 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 co- investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the US and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55–90, to participate in the research, approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years and 200 people with early AD to be followed for 2 years. For up-to-date information, see www.adni-info.org.

Appendix B. ADNI data used in Section 3.5

The following scans from the ADNI database were used in Section 3.5: 109935125-005-1 160830125-005-1 174105125-005-1 51435125-005-1 1425125-005-1 68085125-005-1 77310125-005-1 92610125-005-1 126180125-005-1 166365125-005-1 186705125-005-1 52605125-005-1 63945125-005-1 70380125-005-1 80190125-005-1 98280125-005-1 150075125-005-1 168300125-005-1 46260125-005-1 55575125-005-1 64485125-005-1 73665125-005-1 82530125-005-1 151830125-005-1 170460125-005-1 49095125-005-1 60705125-005-1 66060125-005-1 74745125-005-1 85185125-005-1.

References

Aljabar, P., Heckemann, R.A., Hammers, A., Hajnal, J.V., Rueckert, D., 2009. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. NeuroImage 46 (3), 726–738.

Artaechevarria, X., Munoz-Barrutia, A., Ortiz-de Solorzano, C., 2009. Combination strategies in multi-atlas image segmentation: application to brain MR data. IEEE Transactions on Medical Imaging 28 (8), 1266–1277.

Asman, A.J., Landman, B.A., 2011. Robust statistical label fusion through consensus level, labeler accuracy, and truth estimation (COLLATE). IEEE Transactions on Medical Imaging 30 (10), 1779–1794.

Asman, A.J., Landman, B.A., 2012. Formulating spatially varying performance in the statistical fusion framework. IEEE Transactions on Medical Imaging 31 (6), 1326–1336.

Barnes, J., Foster, J., Boyes, R.G., Pepple, T., Moore, E.K., Schott, J.M., Frost, C., Scahill, R.I., Fox, N.C., 2008. A comparison of methods for the automated calculation of volumes and atrophy rates in the hippocampus. NeuroImage 40 (4), 1655–1671.

Barnes, J., Bartlett, J.W., van de Pol, L.A., Loy, C.T., Scahill, R.I., Frost, C., Thompson, P.M., Fox, N.C., 2009. A meta-analysis of hippocampal atrophy rates in Alzheimer's disease. Neurobiology of Aging 30 (11), 1711–1723.

Cachier, P., Bardinet, E., Dormont, D., Pennec, X., Ayache, N., 2003. Iconic feature based nonrigid registration: the PASHA algorithm. Computer Vision and Image Understanding 89 (2–3), 272–298.

Cardoso, M.J., Leung, K.K., Modat, M., Barnes, J., Ourselin, S., 2011. Locally Ranked STAPLE for Template based Segmentation Propagation, MICCAI Workshop on Multi-Atlas Labeling and Statistical Fusion.

Cardoso, M.J., Clarkson, M.J., Ridgway, G.R., Modat, M., Fox, N.C., Ourselin, S., 2011a. The Alzheimer's disease neuroimaging initiative, LoAd: a locally adaptive cortical segmentation algorithm. NeuroImage 56 (3), 1386–1397.

Cardoso, M.J., Wolz, R., Modat, M., Fox, N., Rueckert, D., Ourselin, S., 2012. Geodesic information flows. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012. Springer Berlin/Heidelberg, Berlin, Heidelberg, pp. 262–270.

Collins, D.L., Pruessner, J.C., 2010. Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion. NeuroImage 52 (4), 1355–1366.

Commowick, O., Akhondi-Asl, A., Warfield, S.K., 2012. Estimating a reference standard segmentation with spatially varying performance parameters: local MAP STAPLE. IEEE Transactions on Medical Imaging 31 (8), 1593–1606.

Dubois, B., Feldman, H.H., Jacova, C., Cummings, J.L., DeKosky, S., Barberger-Gateau, P., Delacourte, A., Frisoni, G.B., Fox, N.C., Galasko, D., Gauthier, S., Hampel, H., Jicha, G.A., Meguro, K., O'Brien, J., Pasquier, F., Robert, P., Rossor, M.N., Salloway, S., Sarazin, M., de Souza, L.C., Stern, Y., Visser, P.J., Scheltens, P., 2010. Revising the definition of Alzheimer's disease: a new lexicon. Lancet Neurol 9 (11), 1118–1127.

Frisoni, G.B., Jack, C.R., 2011. Harmonization of magnetic resonance-based manual hippocampal segmentation: a mandatory step for wide clinical use. Alzheimer's and Dementia 7 (2), 171–174.

Hammers, A., Allom, R., Koepp, M.J., Free, S.L., Myers, R., Lemieux, L., Mitchell, T.N., Brooks, D.J., Duncan, J.S., 2003. Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. Human Brain Mapping 19 (4), 224–247.

Hammers, A., Chen, C.-H., Lemieux, L., Allom, R., Vossos, S., Free, S.L., Myers, R., Brooks, D.J., Duncan, J.S., Koepp, M.J., 2007. Statistical neuroanatomy of the human inferior frontal gyrus and probabilistic atlas in a standard stereotaxic space. Human Brain Mapping 28 (1), 34–48.

Heckemann, R.A., Keihaninejad, S., Aljabar, P., Rueckert, D., Hajnal, J.V., Hammers, A., 2010. Alzheimer's disease neuroimaging initiative, improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation. NeuroImage 51 (1), 221–227.

Henneman, W.J.P., Sluimer, J.D., Barnes, J., van der Flier, W.M., Sluimer, I.C., Fox, N.C., Scheltens, P., Vrenken, H., Barkhof, F., 2009. Hippocampal atrophy rates in Alzheimer disease: added value over whole brain volume measures. Neurology 72 (11), 999–1007.

Jack, C.R., Petersen, R.C., Xu, Y.C., O'Brien, P.C., Smith, G.E., Ivnik, R.J., Boeve, B.F., Waring, S.C., Tangalos, E.G., Kokmen, E., 1999. Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. Neurology 52 (7), 1397–1403.

Jack, C.R., Petersen, R.C., Xu, Y.C., O'Brien, P.C., Smith, G.E., Ivnik, R.J., Boeve, B.F., Tangalos, E.G., Kokmen, E., 2000. Rates of hippocampal atrophy correlate with change in clinical status in aging and AD. Neurology 55 (4), 484–489.

Jack, C.R., Barkhof, F., Bernstein, M.A., Cantillon, M., Cole, P.E., DeCarli, C., Dubois, B., Duchesne, S., Fox, N.C., Frisoni, G.B., Hampel, H., Hill, D.L.G., Johnson, K., Mangin, J.-F.F., Scheltens, P., Schwarz, A.J., Sperling, R.A., Suhy, J., Thompson, P.M., Weiner, M.W., Foster, N.L., 2011. Steps to standardization and validation of hippocampal volumetry as a biomarker in clinical trials and diagnostic criterion for Alzheimer's disease. Alzheimer's and Dementia 7 (4), 474–485.

Lam, L., Suen, C.Y., 1995. Optimal combinations of pattern classifiers. Pattern Recognition Letters 16 (9), 945–954.

Leung, K.K., Barnes, J., Ridgway, G.R., Bartlett, J.W., Clarkson, M.J., Macdonald, K., Schuff, N., Fox, N.C., Ourselin, S., 2010. Alzheimer's disease neuroimaging initiative, automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. NeuroImage 51 (4), 1345–1359.

Lötjönen, J., Wolz, R., Koikkalainen, J., Julkunen, V., Thurfjell, L., Lundqvist, R., Waldemar, G., Soininen, H., Rueckert, D., 2011. Alzheimer's disease neuroimaging initiative, fast and robust extraction of hippocampus from MR images for diagnostics of Alzheimer's disease. NeuroImage 56 (1), 185–196.

Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S., 2010. Fast free-form deformation using graphics processing units. Computer Methods and Programs in Biomedicine 98 (3), 278–284.

Morra, J.H., Tu, Z., Apostolova, L.G., Green, A.E., Avedissian, C., Madsen, S.K., Parikshak, N., Toga, A.W., Jack, C.R., Schuff, N., Weiner, M.W., Thompson, P.M., 2009. Alzheimer's disease neuroimaging initiative, automated mapping of hippocampal atrophy in 1-year repeat MRI data from 490 subjects with Alzheimer's disease, mild cognitive impairment, and elderly controls. NeuroImage 45 (1 Suppl.), S3–15.

Ourselin, S., Roche, A., Prima, S., Ayache, N., 2000. Block matching: a general framework to improve robustness of rigid registration of medical images. In: G. Goos, J. Hartmanis, J. Leeuwen, S.L. Delp, A.M. DiGoia, B. Jaramaz (Eds.), Medical Image Computing and Computer-Assisted Intervention, MICCAI 2000, pp. 557–566.

Ourselin, S., Roche, A., Subsol, G., Pennec, X., 2001. Reconstructing a 3D structure from serial histological sections. Image and Vision Computing 19, 25–31.

Ridha, B.H., Barnes, J., van de Pol, L.A., Schott, J.M., Boyes, R.G., Siddique, M.M., Rossor, M.N., Scheltens, P., Fox, N.C., 2007. Application of automated medial temporal lobe atrophy scale to Alzheimer disease. Archives of Neurology 64 (6), 849–854.

Rohlfing, T., Russakoff, D.B., Maurer Jr, C.R., 2004. Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. IEEE Transactions on Medical Imaging 23 (8), 983–994.

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

Schuff, N., Woerner, N., Boreta, L., Kornfield, T., Shaw, L.M., Trojanowski, J.Q., Thompson, P.M., Jack, C.R., Weiner, M.W., 2009. Alzheimer's disease neuroimaging initiative, MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. Brain 132 (Pt 4), 1067–1077.

Segonne, F., Dale, A.M., Busa, E., Glessner, M., Salat, D.H., Hahn, H., Fischl, B., 2004. A hybrid approach to the skull stripping problem in MRI. NeuroImage 22 (3), 1060–1075.

Souvenir, R., Pless, R., 2007. Image distance functions for manifold learning. Image and Vision Computing 25 (3), 365–373.

Sperling, R.A., Aisen, P.S., Beckett, L.A., Bennett, D.A., Craft, S., Fagan, A.M., Iwatsubo, T., Jack, C.R., Kaye, J., Montine, T.J., Park, D.C., Reiman, E.M., Rowe, C.C., Siemers, E., Stern, Y., Yaffe, K., Carrillo, M.C., Thies, B., Morrison-Bogorad, M., Wagster, M.V., Phelps, C.H., 2011. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's and Dementia 7 (3), 280–292.

Warfield, S.K., Zou, K.H., Wells III, W.M., 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging 23 (7), 903–921.

Wolz, R., Aljabar, P., Hajnal, J.V., Hammers, A., Rueckert, D., 2010. Alzheimer's disease neuroimaging initiative, LEAP: learning embeddings for atlas propagation. NeuroImage 49 (2), 1316–1325.

Woods, K., Kegelmeyer, W.P., Bowyer, K., 1997. Combination of multiple classifiers using local accuracy estimates. IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (4), 405–410.

Xu, L., Krzyzak, A., Suen, C.Y., 1992. Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Transactions on Systems Man and Cybernets 22 (3), 418–435.

Yushkevich, P.A., Wang, H., Pluta, J., Das, S.R., Craige, C., Avants, B.B., Weiner, M.W., Mueller, S., 2010. Nearly automatic segmentation of hippocampal subfields in in vivo focal T2-weighted MRI. NeuroImage 53 (4), 1208–1224.