Medical Image Analysis 42 (2017) 274–287

Contents lists available at ScienceDirect

Medical Image Analysis

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

Discriminative confidence estimation for probabilistic multi-atlas label fusion

Oualid M. Benkarim[a,*], Gemma Piella[a], Miguel Angel González Ballester[a,b], Gerard Sanroma[a], for the Alzheimer's Disease Neuroimaging Initiative[1]

[a] Universitat Pompeu Fabra, Barcelona, Spain
[b] ICREA, Barcelona, Spain

ARTICLE INFO

Article history:
Received 31 January 2017
Revised 26 June 2017
Accepted 29 August 2017
Available online 1 September 2017

Keywords:
Multi-atlas segmentation
Confidence estimation
Discriminative learning
Brain MRI

ABSTRACT

Quantitative neuroimaging analyses often rely on the accurate segmentation of anatomical brain structures. In contrast to manual segmentation, automatic methods offer reproducible outputs and provide scalability to study large databases. Among existing approaches, multi-atlas segmentation has recently shown to yield state-of-the-art performance in automatic segmentation of brain images. It consists in propagating the labelmaps from a set of atlases to the anatomy of a target image using image registration, and then fusing these multiple warped labelmaps into a consensus segmentation on the target image. Accurately estimating the contribution of each atlas labelmap to the final segmentation is a critical step for the success of multi-atlas segmentation. Common approaches to label fusion either rely on local patch similarity, probabilistic statistical frameworks or a combination of both. In this work, we propose a probabilistic label fusion framework based on atlas label confidences computed at each voxel of the structure of interest. Maximum likelihood atlas confidences are estimated using a supervised approach, explicitly modeling the relationship between local image appearances and segmentation errors produced by each of the atlases. We evaluate different spatial pooling strategies for modeling local segmentation errors. We also present a novel type of label-dependent appearance features based on atlas labelmaps that are used during confidence estimation to increase the accuracy of our label fusion. Our approach is evaluated on the segmentation of seven subcortical brain structures from the MICCAI 2013 SATA Challenge dataset and the hippocampi from the ADNI dataset. Overall, our results indicate that the proposed label fusion framework achieves superior performance to state-of-the-art approaches in the majority of the evaluated brain structures and shows more robustness to registration errors.

1. Introduction

Brain segmentation from magnetic resonance imaging (MRI) is an important preprocessing step for many neuroimaging studies, e.g., volumetry, cortical thickness, *etc*. For this task, automatic methods are desirable over manual segmentation since the latter is very time-consuming and subject to inter- and intra-rater variability. Although good outcomes can be achieved for the segmenation of the main tissues based only on image intensities (Leemput et al., 1999; Ashburner and Friston, 2005; Shattuck et al., 2001), segmentation of anatomical structures (e.g., defined by their functional properties) renders intensity information insufficient and atlas priors become an imperative resource in order to accurately delineate such structures. In this setting, single-atlas based segmentation uses a single atlas that is registered to the to-be-segmented image and then propagates its labelmap to the target using the resulting warp from the registration step. Single-atlas based segmentation, however, suffers from (1) representative bias in that a single atlas may not capture the neuroanatomical variability of the general population, and (2) high sensitivity to registration errors since only one atlas is used. To address these drawbacks, multi-atlas segmentation (MAS) makes use of multiple atlases to segment a given target image (Aljabar et al., 2009; Heckemann et al., 2006; Lötjönen et al., 2010). In this way, it better adapts to the anatomical variability of the population and highly mitigates the effect of registration failures in the final segmentation.

* Corresponding author.
E-mail address: oualid.benkarim@upf.edu (O.M. Benkarim).

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

http://dx.doi.org/10.1016/j.media.2017.08.008
1361-8415/

Indeed, MAS has recently shown to be a promising technique for brain structural segmentation (Iglesias and Sabuncu, 2015; Sanroma et al., 2016b). It consists in fusing the propagated labelmaps from a set of training atlases to a target image. There are two main steps: 1) image registration, where the spatial transformations are computed to warp the atlas labelmaps to the target image, and 2) label fusion, where these candidate segmentations (i.e., warped labelmaps) are fused into a consensus segmentation. The focus of this paper is on improving label fusion, the second step of MAS. Label fusion is a rather challenging problem that consists in finding the optimal combination of the propagated atlas labelmaps at each region of the target image to obtain the best segmentation. The most straightforward way to approach this problem is to use majority voting (MV) (Rohlfing et al., 2004; Klein et al., 2005), which assigns to each target the most frequent label occurring among the training atlases. This method has shown superior performance over single-atlas based label propagation. However, since all the atlases are combined with equal weight, having atlases too dissimilar to the target will push the resulting segmentation away from the true target anatomy. In order to solve this problem, several works have proposed more robust label fusion strategies that weigh each atlas vote contribution based on its similarity to the target image (e.g., Artaechevarria et al., 2009; Sabuncu et al., 2010; Coupé et al., 2011). STAPLE (Warfield et al., 2004) and similar methods use a statistical approach to label fusion. Although STAPLE was initially conceived to globally assess the performance of different raters, many works build on STAPLE to provide spatially varying statistical label fusion approaches for MAS. Non-Local STAPLE (Asman and Landman, 2013) and STEPS (Cardoso et al., 2013) extend STAPLE by including appearance information from the images and integrating the non-local means approach (Buades et al., 2005) into the statistical framework of STAPLE. Other works tackle label fusion in MAS from the machine learning perspective, using classification-based approaches (Powell et al., 2008; Sdika, 2015; Zikic et al., 2013), reconstruction-based approaches (Zhang et al., 2012; Benkarim et al., 2014), or a combination of both (Sanroma et al., 2015; 2016a).

In this work, we propose a probabilistic framework with the following contributions:

- We estimate spatially varying confidences for each training atlas in an offline way to reduce computation burden at test time.
- We formulate our method in a probabilistic framework and obtain maximum likelihood confidence parameters through discriminative learning.
- We explore different spatial pooling strategies for modeling local segmentation errors.
- We propose novel label-dependent features to be used together with appearance features to estimate the confidences in the proposed framework.

This paper is an extension of a recently published conference paper (Benkarim et al., 2016). In this current work, we implement more sophisticated spatial pooling strategies to make our method more accurate and computationally efficient, present a more extensive description of the proposed label fusion framework, evaluate the performance of our approach in 2 brain MRI datasets, assess the robustness of our method to registration failures by using several registration settings (affine and 2 different non-rigid registrations), provide an in-depth review of the literature and a comparison of our approach with the state-of-the-art, and include a thorough discussion of the results.

The outline of the paper is as follows. Section 2 is devoted to state-of-the-art label fusion approaches in MAS. Section 3 presents the details of our proposed method. In Section 4 we describe the experimental setting and present the results. In Section 5 we discuss the advantages and limitations of our approach. Section 6 concludes the paper.

2. Related work

The selection of the label fusion strategy is a crucial step in MAS and has been extensively studied in the literature. Label fusion approaches can be grouped in 3 categories according to the strategy used for fusing the different atlas labelmaps to produce the final segmentation: similarity-based, statistical-based and learning-based approaches (Iglesias and Sabuncu, 2015; González-Villà et al., 2016).

2.1. Similarity-based approaches

One major trend is to assign weights (or confidences) to each warped atlas labelmap based on the similarity of its intensity image with the target image. Label fusion with MV can be viewed as a trivial case of these approaches with atlas labelmaps combined with uniform weights. The main assumption of similarity-based approaches is that regions with similar intensities have similar labeling. Several works used this heuristic to perform label fusion. Global weighted voting assigns to each registered atlas a global weight based on its overall similarity with the target image (Artaechevarria et al., 2009). This approach, however, does not consider the spatially varying accuracy of registration, and subsequently, of the confidences. Among approaches that tackle this problem, we can find works using local (Artaechevarria et al., 2009; Isgum et al., 2009; Sabuncu et al., 2010) and non-local (Coupé et al., 2011; Rousseau et al., 2011) weighted voting. Local weighted voting uses one-to-one correspondences of the atlases and the target image, which compensates for the potential misalignments by increasing the weights of the locally well-aligned atlases (and reducing the weights of the rest). In non-local weighted voting, the one-to-one correspondence constraint is relaxed by adopting the non-local means approach proposed in Buades et al. (2005), offering even more flexibility to compensate for registration errors. Moreover, coarsely warped atlases are enough to achieve satisfactory results, thus leading to moderate computational requirements during image registration. In this last approach, the confidence of each atlas is measured using the most similar or all patches from a small search neighborhood around a given voxel. This non-local patch-based strategy has been widely adopted by subsequent methods. The method proposed in Wang et al. (2013), for example, searches for the most similar patch from each atlas and models pairwise dependencies between atlases to reduce the weights of correlated atlases during label fusion.

In similarity-based approaches, performance is sensitive to the choice of the similarity measure, and more importantly, image similarity does not always correlate well with atlas confidence (Sanroma et al., 2014).

2.2. Statistical approaches

Another kind of weighting schemes alleviate the bias induced by similarity-based label fusion by estimating atlas confidences through a more direct measure of the anatomical overlap (Warfield et al., 2004). These approaches alternate the segmentation of the target anatomy and confidence estimation for each of the competing candidate labelmaps by comparing to a consensus segmentation in an iterative fashion. STAPLE is the most representative work, and defines a principled statistical framework based on the Expectation-Maximization (EM) algorithm to perform such estimation. STAPLE, however, was initially conceived to assess the performance of different raters and its performance in MAS is

not significantly better than MV (Artaechevarria et al., 2009; Asman and Landman, 2013). Furthermore, STAPLE does not take into consideration the intensity information available from the images during the confidence estimation process. Many extensions build on STAPLE to provide statistical label fusion approaches for MAS. Gorthi et al. (2014) proposed an approach that incorporates the versatility of local similarity-based approaches into the estimation of the confidences. STEPS (Cardoso et al., 2013) proposes a local ranking strategy based on image similarity to improve the confidence estimation in STAPLE on a voxel-by-voxel basis. The Non-Local STAPLE (Asman and Landman, 2013) integrates the non-local means approach (Buades et al., 2005) and includes appearance information into the statistical framework of STAPLE. Nonetheless, as pointed out earlier, using image similarity can induce a bias in the estimation of the confidence.

2.3. Learning-based approaches

Learning-based methods constitute a different approach to MAS. They attempt to learn, from a set of examples extracted from the training atlases, a function that maps local image appearances to the correct label. A global classifier per atlas using Random Forest (Breiman, 2001) was proposed in Morra et al. (2010) and Zikic et al. (2013). Compared to patch-based approaches, the use of global classifiers further relaxes the one-to-one correspondence constraint. However, global classifiers are usually limited in capturing the complex appearance patterns associated with structural segmentation. This can be circumvented to some extent by using region- or structure-wise classifiers (Powell et al., 2008; Wang et al., 2011), and/or the feature vectors can be augmented to include spatial information for the classifier. Voxel-wise classifiers were also used in the literature for MAS. The work presented in Hao et al. (2014) proposed a MAS approach to estimate the target image's label that learns voxel-wise support vector machine (SVM) classifiers based on the voxel's k nearest positive and negative training samples. Sdika (2015) also used voxel-wise SVM classifiers in a single-atlas based segmentation framework. This approach can be extended to the MAS framework by learning such classifiers in each of the atlas spaces.

The advantage of supervised approaches to MAS is that they can incorporate additional features (e.g., texture, shape, spatial location, *etc.*) (Hao et al., 2014; Bai et al., 2015), which may benefit the classifiers. Furthermore, learning can be performed offline (Zikic et al., 2013; Sdika, 2015), reducing the computational burden of training the classifiers for each target image.

3. Methodology

In this section, we provide a description of the proposed probabilistic label fusion framework. Fig. 1 shows the pipeline of the method, which is composed of two phases:

- Training phase: for each atlas, we compute its confidence model by maximum likelihood estimation. For this task, registration of each atlas to the spaces of the remaining training atlases is first carried out. Then, confidence models in each atlas space are estimated in an offline manner. We propose two ways of estimating the confidence models: 1) a naive approach depending only on local label statistics, and 2) a learning-based approach modeling the relationship between local image appearances and segmentation errors. Confidence estimation in the space of each atlas is important in order to cope with systematic segmentation errors caused by registration failures.
- Testing phase: for a given target image, *spatial confidence maps* (SCMs) are obtained after supplying the target image to the confidence models computed in the training phase. The target's final segmentation is then estimated in a voxel-by-voxel basis with the proposed framework using the SCMs in conjunction with the atlas labelmaps.

3.1. Probabilistic label fusion

In the MAS setting, we have a set of atlas images $\mathbf{A}$ along with their labelmaps $\mathbf{D}$, where $D_{ij} \in \mathbf{D}$ and $D_{ij} = \{1, \ldots, p\}$, indicates which one of the p structures is present at voxel i of the jth atlas. Now consider a novel target image T, where T_i denotes the intensity value at voxel i, we denote the to-be-estimated target labelmap as F.

Our proposed label fusion follows the derivation of a spatially varying version of STAPLE proposed in Asman and Landman (2012). The goal is to find the target labels that maximize the following posterior probability:

$$f(F|\mathbf{D}, \mathbf{C}) = \prod_i f(F_i|\mathbf{D}_i, \mathbf{C}_i) = \prod_i \frac{f(\mathbf{D}_i|F_i, \mathbf{C}_i) f(F_i)}{f(\mathbf{D}_i, |\mathbf{C}_i)}, \tag{1}$$

where $\mathbf{D}_i$ denotes the set of atlas decisions for voxel i and $\mathbf{C}_i$ denotes their respective confidences (or weights). Note that we assume conditional independence in the target voxels. Further assuming independence among the atlas decisions, we obtain the following expression:

$$f(F_i|\mathbf{D}_i, \mathbf{C}_i) = \frac{\prod_j f\left(D_{ij}|F_i, C_{ij}\right) f(F_i)}{\sum_{s\in\{1,0\}} \prod_j f\left(D_{ij}, |F_i = s, C_{ij}\right) f(F_i = s)}. \tag{2}$$

The binary segmentation case is considered in Eq. (2), i.e., we have only two labels denoted {0, 1}. For multiple structures, a one-versus-rest approach can be used.

Accordingly, the probability of the target label F_i being foreground (i.e., label 1) is defined as:

$$f(F_i = 1|\mathbf{D}_i, \mathbf{C}_i) = \frac{a_i}{a_i + b_i}, \tag{3}$$

where

$$a_i = f(F_i = 1) \prod_j f\left(D_{ij}|F_i = 1, C_{ij}\right) \tag{4}$$

$$b_i = f(F_i = 0) \prod_j f\left(D_{ij}|F_i = 0, C_{ij}\right). \tag{5}$$

Here, we are interested in $f\left(D_{ij}|F_i = s, C_{ij}\right)$, which is the probability of observing the decision of jth atlas on voxel i, given that the target label is s and the atlas confidence at that voxel is C_{ij}. This term expresses the likelihood that the atlas and target labels coincide, and is defined as:

$$f\left(D_{ij}|F_i = s, C_{ij}\right) = \begin{cases} C_{ij} & \text{if } D_{ij} = s \\ 1 - C_{ij} & \text{otherwise.} \end{cases} \tag{6}$$

In the EM framework used by STAPLE-based approaches, Eq. (3) corresponds to the estimation of the hidden reference segmentation (i.e., E-step) given the rater performance parameters or confidences, C_{ij}. These confidences are then updated during the M-step based on the previous E-step, and this process is repeated interleaving both steps until reaching convergence. The main difference of our approach with Asman and Landman (2012) lies in the computation of the C_{ij} confidences in Eq. (6), which is the central part of our work. We propose to estimate spatially varying confidences (i.e., for each voxel) in an offline manner using the atlases in the training set instead of the iterative EM-based approach used in Asman and Landman (2012).

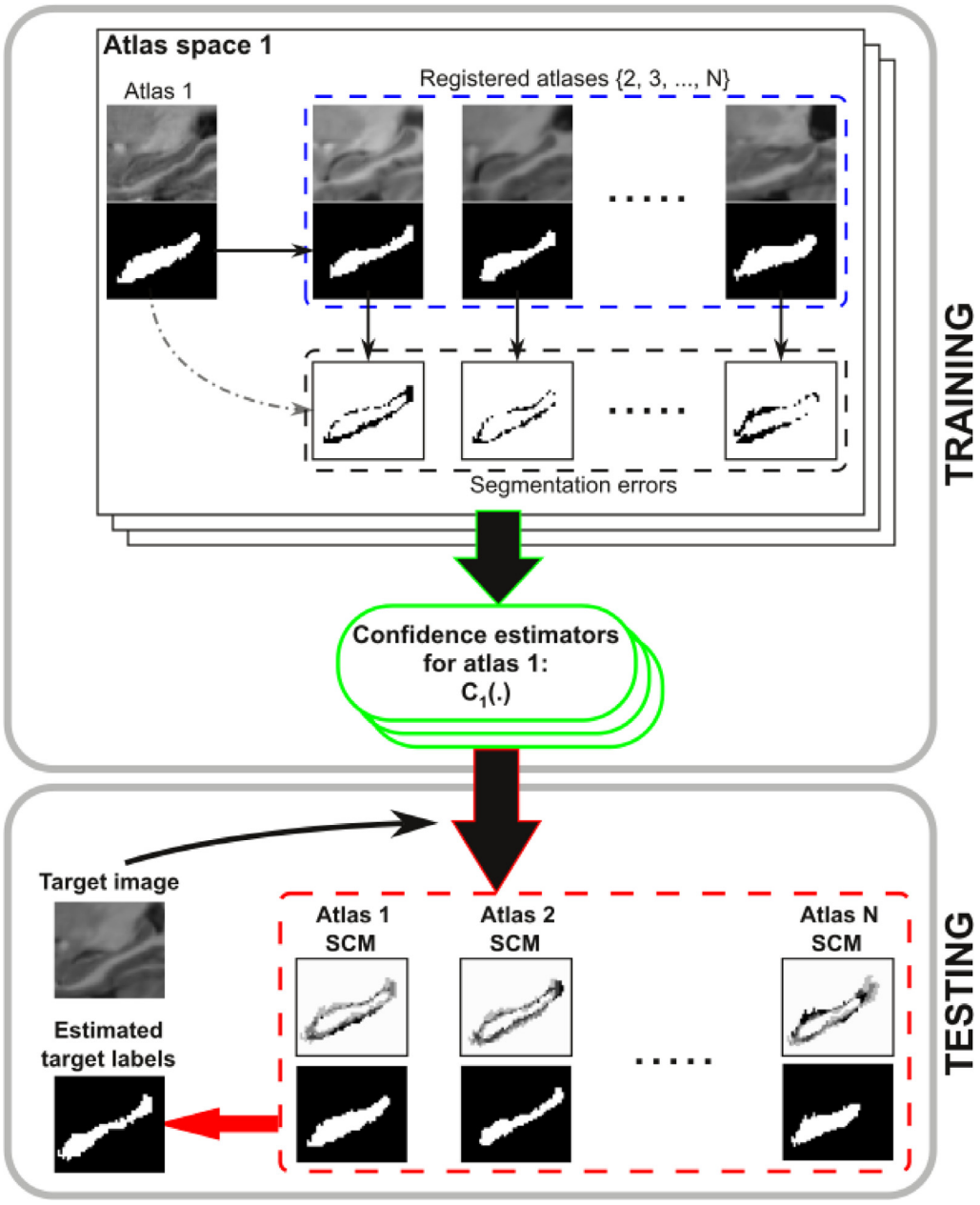

Fig. 1. Pipeline of the proposed label fusion approach. **Training**: for each atlas, the remaining atlases are registered onto the atlas space and confidence models are computed. **Testing**: given a novel to-be-segmented image, SCMs from each atlas are obtained using the confidence models. Target labels are then estimated according to the proposed label fusion framework. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.2. Confidence estimation

Let us focus on the computation of the confidence for a single voxel i of a single atlas j, denoted as $c \equiv C_{ij}$ for brevity (the same procedure is repeated for the rest of the voxels on the rest of atlases). Similarly, let us denote as $d \equiv D_{ij}$ the label at voxel i in the jth atlas. We denote as $\mathbf{f} = \left\{\tilde{D}_{ik}, k \neq j\right\}$, the training set of target observations for the voxel i in the jth atlas composed of the registered labelmaps of the rest of atlases. This is indicated by the blue panel in Fig. 1. We compute the confidence at each voxel by maximizing the following joint likelihood:

$$\begin{aligned}\hat{c} &= \arg\max_c f(\mathbf{f}, d|c) \\ &= \arg\max_c \prod_k f(d|f_k, c) f(f_k|c),\end{aligned} \tag{7}$$

where $f_k \in \mathbf{f}$. We discard the second term in the product since we assume that target labels are only affected by the confidence parameters in the presence of an atlas. Taking the logarithm and substituting the atlas likelihood term by its expression in Eq. (6) yields:

$$\begin{aligned}\hat{c} &= \arg\max_c \sum_k \log f(d|f_k, c) \\ &= \arg\max_c \sum_{f_k = d} \log c + \sum_{f_k \neq d} \log(1 - c).\end{aligned} \tag{8}$$

Taking derivatives, the optimal confidence is

$$c = \frac{n_h}{n_h + n_m}, \tag{9}$$

where n_h and n_m are the number of coincident target labels (*hits*) and different target labels (*misses*), respectively, from the atlas la-

bel. This defines our naive approach. When all atlases are used to compute the confidences, this approach yields similar results to MV. Note that Eq. (7) is the analogue of the M-step in STAPLE-based approaches. However, we are using solely the training atlases and no estimation of the true hidden segmentation is considered, as opposed to Asman and Landman (2012).

Nevertheless, we further believe that local image appearances provide valuable clues for estimating this confidence. Therefore, we extend the previous naive method by substituting the constant confidence in Eq. (6) by a more complex function informed by the image appearances, as follows:

$$f\left(D_{ij}|F_i = s, \mathcal{C}_{ij}\right) = \begin{cases} \mathcal{C}_{ij}\left(\mathbf{t}_i, \mathbf{a}_{ij}\right) & \text{if } D_{ij} = s \\ 1 - \mathcal{C}_{ij}\left(\mathbf{t}_i, \mathbf{a}_{ij}\right) & \text{otherwise} \end{cases}, \tag{10}$$

where $\mathbf{t}_i$ and $\mathbf{a}_{ij}$ are image appearance features extracted around voxel i from the target atlas image and the jth atlas respectively. $\mathcal{C}_{ij}(\cdot)$ is a function denoting the confidence we have that the atlas label is correct given the target and atlas image appearances (as shown in the green panel in Fig. 1). By using image appearances, we can effectively capture the effects of registration errors on modeling such confidence. Again, our goal is to compute such function as to maximize the joint probability of each atlas observation given the training set. Using a similar development as in the naive case, we arrive at the following expression:

$$\hat{\mathcal{C}} = \arg\max_{\mathcal{C}} \sum_{f_k = d} \mathcal{C}(\mathbf{t}_k, \mathbf{a}) - \sum_{f_k \neq d} \mathcal{C}(\mathbf{t}_k, \mathbf{a}), \tag{11}$$

where $\mathbf{t}_k$ and $\mathbf{a}$ denote the local image appearances of the kth target training sample and atlas in the training set, respectively.

In the testing stage, given a new target image T, it is first warped to each of the atlases in the training set. Then, SCMs are computed using the confidence functions of Eq. (11) based on intensity information from both the target, T, and the atlases, $\mathbf{A}$. Next, SCMs and their corresponding atlas labelmaps, $\mathbf{D}$, are transformed back to the target space. Finally, we compute the label fusion using Eq. (1), as shown in the red panel of Fig. 1.

3.2.1. Training

Expression (11) corresponds to the minimization of an empirical error subject to the constraint that the computed function must be a probability density function. For this purpose, we consider a learning-based approach to build voxel-wise classifiers as our confidence estimators. Note that we segment each structure separately, thus using binary classifiers. In order to explain the procedure to create the samples used to train each voxel-wise classifier, let us assume the simple case of one-to-one correspondences. For each training atlas (in its native space), classifiers are built for each voxel. Consider an atlas $A \in \mathbf{A}$ in its native space and a target atlas $W \in \mathbf{A} \backslash \{A\}$ warped to A. For the ith voxel, let a_i and w_i respectively represent the patches of atlas A and the warped target atlas W, with corresponding labels d_i^a and d_i^w. That is, the pair:

1. (a_i, d_i^a) represents the patch and label of the ith voxel in atlas A.
2. (w_i, d_i^w) represents the patch and label of the ith voxel in the target atlas W.

These 2 pairs are used to create a single training sample (x_i, y_i) corresponding to atlas W for the ith classifier of atlas A as follows:

- For the features, we use a patch-based approach. The feature vector, x_i, consists on the intensity difference between the atlas patch and the patch from the target:

  $$x_i = a_i - w_i.$$

- The class label (i.e., the label used to train our classifiers), y_i, is built from the atlas labels (i.e., the voxel labels, d_i^a and d_i^w) and corresponds to the segmentation error produced by the atlas (i.e., A) when segmenting the target (i.e., W), and is defined as:

  $$y_i = \delta(d_i^a, d_i^w),$$

  where $\delta(\cdot, \cdot)$ is the Kronecker delta function. If d_i^a and d_i^w are equal, then y_i is 1, and 0 otherwise. In other words, the class label (or the ground truth during training) tells us if atlas A correctly segments atlas W ($y_i = 1$) or not ($y_i = 0$) at the ith voxel.

Given N training atlases in our database, in this simple case of one-to-one correspondences, the number of training samples used to train the ith classifier of atlas A is $N-1$, where each sample is built from atlas A and each of the remaining $N-1$ warped atlases W. Therefore, the ith classifier of atlas A attempts to learn from all the x_i what are the patterns of intensity differences that lead atlas A to produce correct or erroneous labels, based on the rest of training atlases.

In the test stage, when a novel to-be-segmented image arrives, it is transformed to the spaces of all training atlases. Given the test image T warped to the space of atlas A and the patch t_i of the test image at the ith voxel. The patch difference a_i - t_i is fed to the classifier, which will predict how likely is the label of atlas A at voxel i (i.e., d_i^a) to be the correct label for t_i. The higher the predicted probability by the classifier, the more likely is the test patch to have a similar label to atlas A. This is what we interpret as confidence in our label fusion. Once this has been done for all N training atlases, we will have N label candidates for t_i along with their predicted confidences.

This is how confidences are estimated with Eq. (11). Note that in the naive case, according to Eq. (9), the optimal confidence is equal to the proportion of correct labels in the training set. We do not use the feature vectors x_i, just generate the class labels, y_i, to compute this confidence, which tells us how good is the label d_i^a of atlas A in segmenting the rest of training atlases W.

Instead of using simple one-to-one correspondences, we adopt two different spatial pooling strategies to build the training set for each voxel-wise classifier: 1) non-local means approach in the target space and 2) non-local means approach in both atlas and target spaces. In the following we describe both approaches in detail.

3.2.2. Non-local means approach in target space

Here we use one-to-many correspondences. Given voxel i in the atlas space A, whose patch a_i is represented as the blue box in Fig. 2, we used its label, d_i^a, to segment all voxels within a neighborhood window in the warped atlas W, $S_w(i)$. This is illustrated in Fig. 2 as a red box in the target atlas. The features and labels for A and a given warped atlas W are extracted as follows:

$$x_j = a_i - w_j, \forall j \in S_w(i),$$

$$y_j = \delta(d_i^a, d_j^w), \forall j \in S_w(i).$$

The advantage of this approach is twofold: 1) there are more samples to learn the voxel-wise classifiers than in the one-to-one correspondences case, and 2) confidence estimators are more robust as they are trained to take into account larger registration errors (i.e., all patches in the neighborhood window of W at the ith voxel). The number of samples created to train the classifier is $(N-1) \times |S_w(i)|$, where $|\cdot|$ denotes the size of the neighborhood window (e.g., for a $3 \times 3 \times 3$ neighborhood window, each warped atlas W contributes with 27 samples.)

3.2.3. Non-local means approach in target and atlas spaces

This is an extension of the previous point to have many-to-many correspondences. Here, instead of using a single voxel i in the atlas space A to segment the target W, we take into consideration all voxels in its neighborhood, $S_a(i)$ (depicted in Fig. 3 as a

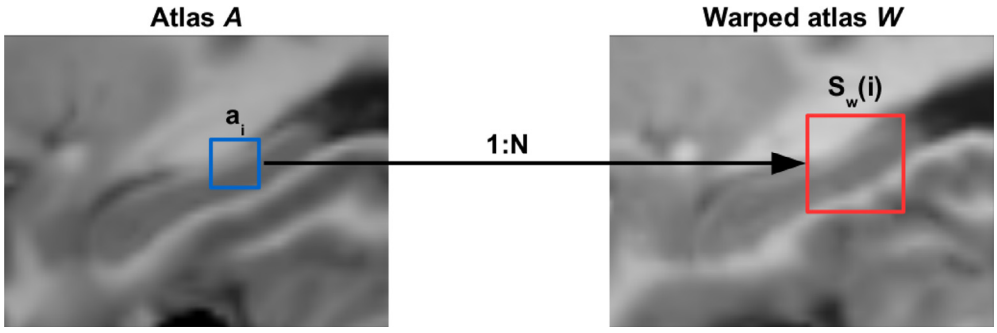

Fig. 2. Non-local means approach in target space. The blue box represents the patch around the *i*th voxel in the atlas space *A*. The red box in the target space *W* represents the window search from which we extract all patches. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

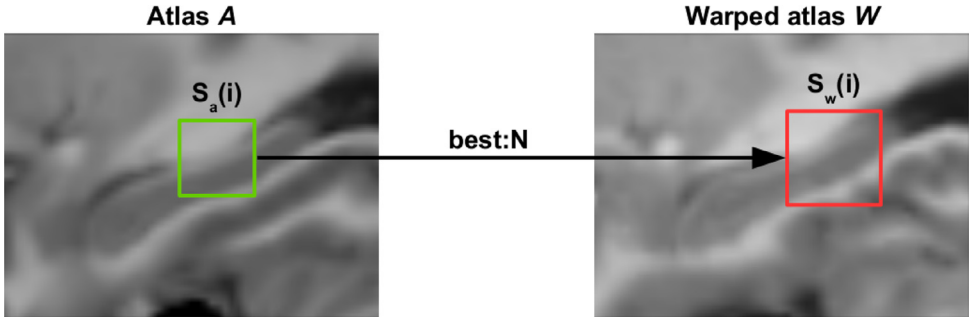

Fig. 3. Non-local means in both spaces. The green box represents the window search from which the best (i.e., most similar) patch is selected for each patch in the target atlas *W* (red box in target atlas). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

green box), and then use k-nearest neighbors to select the voxel label of the most similar atlas patch $a_i^{\hat{k}}$, corresponding to a given target patch, (see arrow labeled *best:N* in Fig. 3). The samples created from *A* and a particular target atlas *W* are defined as:

$$x_j = a_i^{\hat{k}} - w_j, \forall j \in S_w(i),$$

$$y_j = \delta(d_i^{a\hat{k}}, d_j^w), \forall j \in S_w(i),$$

with $\hat{k}$ indexing the atlas patch most similar to w_j:

$$\hat{k} = \underset{k \in S_a(i)}{\operatorname{argmax}}\, sim(a_k, w_j),$$

where *sim* denotes a similarity measure (e.g., cosine similarity). In this way, segmentation errors produced by atlas *A* are based on appearance information. In fact, the voxel-wise classifiers in this case are built upon a similarity-based approach, therefore, learning not only the appearance patterns that lead the current atlas *A* to mislabel the remaining target atlases, but also the behavior of the similarity measure in segmenting the target patches. Furthermore, using many-to-many correspondences without *k*-nearest neighbors to construct the datasets for our confidences estimators will make training computationally expensive given that we learn voxel-wise classifiers. The number of samples used to train the classifiers is the same as in the one-to-many correspondences, being the only difference that the atlas patch $a_i^{\hat{k}}$ in this case is not fixed but the most similar among all patches in $S_a(i)$ to the target patch, w_j. At test time and for the *i*th voxel, given a test patch t_i, the most similar patch and its label from atlas *A* to t_i are selected. The classifier then predicts the confidence of this label (i.e., the label corresponding to the most similar patch in *A*) in correctly segmenting t_i based on the patch difference.

3.2.4. Label-dependent feature extraction

In patch-based approaches, the simplest way to represent local features is to use a cubic patch around the voxel of interest, as stated in Section 3.2.1. Here, to fully take advantage of our learning-based confidence estimators, we propose to use additional features based on the atlas labelmap. This contribution uses the label patch of atlas *A* to extract label-dependent features from the warped images *W*. As illustrated in Fig. 4, given the label patch of the atlas *A* around the *i*th voxel, we identify the target voxels corresponding to foreground and background regions (in the case of binary segmentation) and compute different summary statistics. Finally, the difference between foreground and background features is calculated and the resulting features are appended to the intensity patch.

With our label-dependent features, we attempt to characterize the intensity distributions of the target patches according to a given atlas label patch. In principle, it is expected that background voxels would exhibit a different intensity distribution when compared to foreground voxels since they do belong to different structures. Therefore, the more accurate the atlas label patch is in segmenting the target patches, the larger the features difference would be between these regions.

4. Experiments

In this section, we present the evaluation of our proposed approach and provide a comparison of its performance with state-of-the-art MAS methods for the segmentation of seven subcortical brain structures: accumbens, amygdala, caudate, hippocampus, pallidum, putamen and thalamus proper.

4.1. Data and preprocessing

The proposed approach was evaluated on 2 brain MRI datasets:

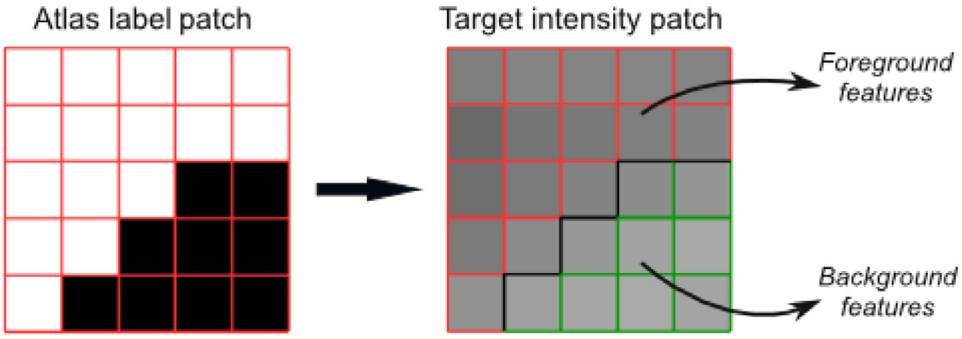

Fig. 4. Label-dependent feature extraction.

1. MICCAI 2013 SATA Challenge dataset[2]: This dataset is composed of 35 T1-weighted MR images of control subjects with age ranging from 19 to 90 years (32.4 years old in average). The size of the images is $256 \times 256 \times 287$ with a spatial resolution of 1 mm isotropic. Ground-truth segmentations are available for all 7 subcortical structures.
2. ADNI dataset[3]: We used a subset of 135 T1-weighted MR images (44 normal controls, 46 subjects with mild cognitive impairment and 45 with Alzheimer's disease). The age distribution is: 40 between 60–70 years, 55 between 70–80 years and 40 with more than 80 years. The size of the images is $197 \times 233 \times 189$ with a voxel size of $1 \times 1 \times 1$ mm. In this dataset, ground-truth segmentations are only available for the hippocampi.

Our method requires pairwise registrations since its needs to have each atlas in the rest of the training spaces. However, to save computational time, all images were registered to a common reference space (i.e., the MNI152 template). Pairwise mappings were then obtained by composing the transformation of the source atlas to the template space and the inverse transformation from the template to the target atlas. Furthermore, for image intensity to be consistent across atlases, all images were normalized using histogram matching (Nyul et al., 2000).

4.2. Experimental setup

We evaluated our method using the following configurations:

- Naive: the naive approach, where segmentation is based only on local label statistics (i.e., voxel-wise label errors as priors).
- SCMNF: the SCM approach using one-to-many correspondences with only patch intensities as features.
- SCMWF: similar to SCMNF but including label-dependent features.
- SCMNF2: the SCM approach based on many-to-many correspondences with only patch intensities as features.
- SCMWF2: similar to SCMNF2 but including label-dependent features.

For comparison, we considered the following state-of-the-art methods: MV, local weighted voting with inverse similarity metric (LWV) (Artaechevarria et al., 2009), STAPLE, STEPS and joint label fusion (JOINT) (Wang et al., 2013).

The summary statistics we used as label-dependent features for SCMWF and SCMWF2 in the experiments were: mean, maximum and minimum intensities, and the center of mass of each region. Regarding the classifiers used for our confidence estimators, we used logistic regression. For SCMNF2 and SCMWF2, the similarity measure used to select the best atlas patch is cosine similarity:

$$cos(x_i, x_j) = \frac{x_i \cdot x_j}{||x_i|| \cdot ||x_j||}.$$

No parameter tuning was performed for the experiments. We used the default values for all methods, except for the radius of the patch and window search that was set to 1 (i.e., a patch and window search size of $3 \times 3 \times 3$). For logistic regression, the penalty parameter C was set to 1.

For the SATA dataset a 3-fold cross-validation procedure was used in our evaluation strategy, and for ADNI, 35 atlases were selected for training and the remaining 100 for test. The 35 training atlases were selected in order to span the space of all images using spectral clustering based on normalized correlation. For quantitative comparison, we used the Dice similarity coefficient (Dice, 1945), determined as follows:

$$D(A, B) = \frac{2|A \cap B|}{|A| + |B|},$$

where A and B are the reference and automatic segmentations, respectively, and the modified Hausdorff distance (MHD) (Dubuisson and Jain, 1994), defined as:

$$MHD(S, T) = \max(d(S, T), d(T, S))$$

where S and T are the sets of voxels in the boundary of A and B respectively, and d is a directed distance measure between the first and the second sets based on Euclidean distance. MHD is reported in mm throughout the whole article. Statistical significance is measured using the Wilcoxon signed rank test and is reported at $p < 0.05$.

Finally, in order to assess the robustness of our approach to registration failures, all experiments were replicated using three different registrations settings:

1. AF: Affine registration,
2. NR1: Affine followed by a non-rigid registration at a coarse scale using the symmetric diffeomorphic mapping (SyN) proposed by Avants et al. (2008). Non-rigid registration was done in a multi-resolution fashion using a regular grid with control point spacings of 8 and 4 mm, and
3. NR2: Affine followed by a finer non-rigid registration (NR2) using SyN. Non-rigid registration was done in a multi-resolution fashion using a regular grid with control point spacings of 8, 4, 2 and 1 mm.

4.3. Implementation and computational complexity

Our method was implemented in Python using the logistic regression Python wrapper provided by *Scikit-learn* (Pedregosa et al., 2011) for the *liblinear* library (Fan et al., 2008). For STAPLE and

[2] https://masi.vuse.vanderbilt.edu/workshop2013
[3] http://adni.loni.usc.edu

Table 1
Mean Dice scores (top entries) and MHD (bottom entries) per structure, averaged left and right. Results obtained using the non-rigid registration NR2. Bold type indicates the best segmentation performance in terms of Dice overlap or MHD. The * symbol indicates statistical significance difference with all remaining methods, and † indicates statistical significance difference with all methods except SCMNF2 or SCMWF2. Abbreviations: accumbens (Acc), amygdala (Amy), caudate (Cau), hippocampus (Hip), pallidum (Pal), putamen (Put) and thalamus proper (Tha).

	SATA							ADNI
	Acc	Amy	Cau	Hip	Pal	Put	Tha	Hip
MV	0.777 ± 0.052	0.799 ± 0.038	0.826 ± 0.096	0.831 ± 0.037	0.882 ± 0.027	0.920 ± 0.019	0.908 ± 0.026	0.767 ± 0.049
Naive	0.779 ± 0.052	0.799 ± 0.038	0.828 ± 0.096	0.830 ± 0.037	0.886 ± 0.027	0.920 ± 0.019	0.912 ± 0.026	0.768 ± 0.049
STAPLE	0.767 ± 0.064	0.797 ± 0.041	0.819 ± 0.103	0.828 ± 0.036	0.877 ± 0.027	0.915 ± 0.018	0.904 ± 0.027	0.768 ± 0.058
STEPS	0.768 ± 0.075	0.797 ± 0.044	0.822 ± 0.105	0.832 ± 0.042	0.882 ± 0.029	0.919 ± 0.018	0.908 ± 0.028	0.799 ± 0.043
LWV	0.784 ± 0.053	0.802 ± 0.037	0.863 ± 0.075	0.843 ± 0.030	0.881 ± 0.027	0.919 ± 0.018	0.914 ± 0.022	0.796 ± 0.045
JOINT	0.799 ± 0.039	0.827 ± 0.024	0.888 ± 0.068	0.871 ± 0.021	**0.888 ± 0.027**	**0.926 ± 0.018***	0.923 ± 0.014	0.860 ± 0.037
SCMNF	0.792 ± 0.049	0.812 ± 0.030	0.902 ± 0.051	0.867 ± 0.016	0.885 ± 0.024	0.923 ± 0.026	0.925 ± 0.011	0.844 ± 0.035
SCMWF	0.805 ± 0.047	0.818 ± 0.033	0.905 ± 0.049	0.871 ± 0.016	0.886 ± 0.026	0.923 ± 0.024	0.924 ± 0.010	0.850 ± 0.039
SCMNF2	0.808 ± 0.047	0.825 ± 0.032	0.906 ± 0.042	0.872 ± 0.018	0.885 ± 0.028	0.922 ± 0.021	0.923 ± 0.013	0.853 ± 0.038
SCMWF2	**0.811 ± 0.044***	**0.830 ± 0.028**	**0.907 ± 0.040†**	**0.877 ± 0.015***	0.886 ± 0.027	0.924 ± 0.019	**0.925 ± 0.011**	**0.866 ± 0.026**
MV	2.906 ± 0.930	2.851 ± 0.608	4.259 ± 1.455	4.876 ± 1.500	2.283 ± 0.381	2.578 ± 0.780	3.341 ± 1.066	4.119 ± 1.039
Naive	2.912 ± 0.930	2.854 ± 0.608	4.263 ± 1.455	4.875 ± 1.500	2.285 ± 0.381	2.575 ± 0.780	3.341 ± 1.066	4.119 ± 1.039
STAPLE	2.931 ± 0.889	2.978 ± 0.607	3.918 ± 1.477	4.840 ± 1.413	2.273 ± 0.378	2.570 ± 0.732	3.461 ± 1.277	3.924 ± 0.933
STEPS	3.054 ± 0.976	2.804 ± 0.642	4.114 ± 1.593	4.780 ± 1.576	2.204 ± 0.442	2.406 ± 0.786	3.306 ± 1.164	3.754 ± 0.993
LWV	2.693 ± 0.811	2.800 ± 0.637	3.817 ± 1.118	4.604 ± 1.363	2.273 ± 0.394	2.573 ± 0.757	3.172 ± 0.818	3.855 ± 0.963
JOINT	2.683 ± 0.848	**2.749 ± 0.551**	4.070 ± 1.764	4.846 ± 1.569	2.347 ± 0.439	2.474 ± 0.900	3.304 ± 1.266	3.467 ± 0.868
SCMNF	2.778 ± 0.750	3.125 ± 0.915	3.355 ± 1.042	4.964 ± 1.550	2.412 ± 0.478	2.699 ± 0.879	3.207 ± 0.866	3.850 ± 0.802
SCMWF	2.569 ± 0.635	3.111 ± 0.968	3.155 ± 0.987	4.624 ± 1.397	2.293 ± 0.474	2.555 ± 0.844	3.102 ± 0.938	3.662 ± 0.945
SCMNF2	**2.348 ± 0.565†**	2.892 ± 0.714	3.092 ± 0.969	4.559 ± 1.539	**2.207 ± 0.447**	**2.373 ± 0.798**	**2.906 ± 0.937†**	3.504 ± 0.793
SCMWF2	2.351 ± 0.557	2.846 ± 0.668	**3.075 ± 0.965†**	**4.474 ± 1.333**	2.239 ± 0.446	2.382 ± 0.775	2.922 ± 0.919	**3.369 ± 0.744**

STEPS, we used the implementations distributed in the NiftySeg[4] software package. For JOINT, the implementation shipped with the ANTs[5] package was used.

Experiments were executed on a PC running 64bit Ubuntu Linux 14.04 LTS with a system configuration Intel(R) Core(TM) i7-4790 CPU (3.60 GHz) × 8 with 32GB of RAM.

Execution times required by our offline learning vary depending on the size of the structure (i.e, number of voxel-wise classifiers) and the use of label-dependent features. To reduce the runtimes, learning was not performed for voxels where all the atlases were in consensus (i.e., same label). Training our confidence estimators for the accumbens, for instance, took around 3 and 12 min for SCMNF2 and SCMWF2, respectively. For SCMNF and SCMWF, learning took approximately 3 and 10 min. At test time, all verseconfsions of our method produced segmentations for the accumbens in less than 2 s, similarly to the rest of the methods, except JOINT that took around 10 s. For one of the largest structures, the hippocampus in the ADNI dataset, segmentation took around 4 s for SCMNF, 6 s for SCMNF2, and 12–15 s for SCMWF and SCMWF2. For MV, the Naive approach and STAPLE, segmentation took around 2 s. Segmentation times for STEPS and LWV were less than 5 s and for JOINT took around 20 s.

4.4. Results

Table 1 shows the average Dice overlap per structure (in both brain MRI databases) achieved by each of the approaches considered in the experiments using NR2 registration (i.e., finer non-rigid registration), with which all methods provided the best segmentations. Per structure performance results for AF and NR1 registrations are included in the supplementary material. The performance of our Naive approach in terms of both Dice overlap and MHD is similar to MV. In fact, when all atlas labelmaps are used to compute the constant confidence in Eq. (6), it is equivalent to MV, being the additional transformations between atlas spaces the only difference. We should expect higher segmentation results when using pairwise registration, because, in the conducted experiments, the image to-be-segmented was only registered to the common space. STAPLE did not show superior performance over our Naive approach and MV in this registration setting. The other STAPLE-based approach used in this comparison (i.e., STEPS) yielded slightly better results than the aforementioned methods, especially in the hippocampi from the ADNI dataset. Nonetheless, this improvement was not consistent since STAPLE provided lower MHD for the amygdala, caudate and thalamus proper when using NR2 registration. Furthermore, STEPS was outperformed by LWV in the segmentation results of all structures except the pallidum and the hippocampus from ADNI.

With the exception of STEPS in the SATA dataset and LWV in the ADNI dataset, we can observe a clear dichotomy in performance across all structures, as reported in Table 1, between approaches ignoring image intensity (i.e., MV, Naive and STAPLE) and the rest of methods using appearance information. We can already see a quantitative increase in Dice overlap and a decrease in MHD with LWV and STEPS. However, JOINT and the four versions of our approach provided the most accurate segmentations, with statistically significant improvement in all structures over MV, Naive, STAPLE and STEPS.

Comparing the different intensity-based configurations of our approach, SCMNF outperformed LWV and STEPS, but the considerable boost in performance was due to the inclusion of the label-dependent features in SCMWF, which reached an overall Dice score and MHD comparable to JOINT (see Table 2). Still, when adopting the many-to-many correspondences to learn the confidence estimators, segmentation results with our novel approaches (i.e., SCMNF2 and SCMWF2) were better than their original analogue versions. In fact, SCMNF2 produced similar Dice overlaps and MHD to SCMWF without using label-dependent features. In SCMWF2, the inclusion of these, has further improved the segmentation results in all structures according to Dice overlap, as shown in Table 1. In terms of MHD, SCMWF2 was outperformed by SCMNF2 in the accumbens, pallidum, putamen and thalamus proper, though the differences were minuscule. Dice overlaps achieved by SCMWF2 were statistically higher than the rest of methods in the accumbens, and hippocampus, and in the caudate results were statistically significant except for SCMNF2. For MHD,

[4] http://cmictig.cs.ucl.ac.uk/wiki/index.php/NiftySeg
[5] https://stnava.github.io/ANTs

Table 2
Overall mean Dice scores (top entries) and MHD (bottom entries) per database for each registration setting. Bold type indicates the best segmentation performance in terms of Dice overlap or MHD. The * symbol indicates statistical significance difference with all remaining methods, and † indicates statistical significance difference with all methods except SCMNF2.

	SATA			ADNI		
	AF	NR1	NR2	AF	NR1	NR2
MV	0.737 ± 0.083	0.808 ± 0.056	0.849 ± 0.042	0.635 ± 0.068	0.693 ± 0.055	0.767 ± 0.049
Naive	0.737 ± 0.083	0.808 ± 0.056	0.850 ± 0.042	0.632 ± 0.068	0.694 ± 0.055	0.768 ± 0.049
STAPLE	0.738 ± 0.084	0.815 ± 0.062	0.844 ± 0.045	0.670 ± 0.072	0.711 ± 0.065	0.768 ± 0.058
STEPS	0.746 ± 0.086	0.827 ± 0.033	0.847 ± 0.049	0.733 ± 0.058	0.763 ± 0.048	0.799 ± 0.043
LWV	0.767 ± 0.078	0.847 ± 0.047	0.858 ± 0.037	0.710 ± 0.065	0.748 ± 0.051	0.796 ± 0.045
JOINT	0.855 ± 0.046	0.872 ± 0.034	0.875 ± 0.030	0.835 ± 0.043	0.853 ± 0.031	0.860 ± 0.037
SCMNF	0.844 ± 0.048	0.866 ± 0.032	0.872 ± 0.030	0.811 ± 0.039	0.833 ± 0.034	0.844 ± 0.035
SCMWF	0.849 ± 0.046	0.869 ± 0.030	0.876 ± 0.029	0.818 ± 0.040	0.838 ± 0.037	0.850 ± 0.039
SCMNF2	0.857 ± 0.042	0.871 ± 0.030	0.877 ± 0.028	0.832 ± 0.039	0.848 ± 0.037	0.853 ± 0.038
SCMWF2	**0.865 ± 0.037***	**0.874 ± 0.029**	**0.880 ± 0.026**	**0.843 ± 0.038***	**0.856 ± 0.036**	**0.866 ± 0.026**
MV	4.286 ± 1.070	3.634 ± 1.057	3.299 ± 0.960	5.494 ± 1.289	4.730 ± 1.095	4.119 ± 1.039
Naive	4.287 ± 1.070	3.635 ± 1.057	3.301 ± 0.960	5.493 ± 1.289	4.730 ± 1.095	4.119 ± 1.039
STAPLE	4.221 ± 1.037	3.496 ± 1.123	3.282 ± 0.968	4.872 ± 1.017	4.331 ± 0.871	3.924 ± 0.933
STEPS	3.930 ± 1.047	3.610 ± 1.043	3.238 ± 1.026	4.527 ± 1.254	4.031 ± 1.039	3.754 ± 0.993
LWV	3.938 ± 1.044	3.391 ± 0.904	3.133 ± 0.842	4.778 ± 1.154	4.250 ± 0.974	3.855 ± 0.963
JOINT	3.560 ± 1.215	3.279 ± 1.135	3.210 ± 1.048	4.122 ± 1.019	3.682 ± 1.065	3.467 ± 0.868
SCMNF	3.660 ± 1.053	3.317 ± 0.906	3.220 ± 0.926	4.445 ± 0.931	4.088 ± 0.857	3.850 ± 0.802
SCMWF	3.621 ± 1.095	3.213 ± 0.955	3.058 ± 0.892	4.487 ± 1.069	3.995 ± 0.976	3.662 ± 0.945
SCMNF2	3.382 ± 1.040	3.001 ± 0.878	2.911 ± 0.853	4.026 ± 0.894	3.640 ± 0.821	3.504 ± 0.793
SCMWF2	**3.312 ± 1.011†**	**2.966 ± 0.857**	**2.898 ± 0.809†**	**3.728 ± 0.850***	**3.605 ± 0.839†**	**3.369 ± 0.744**

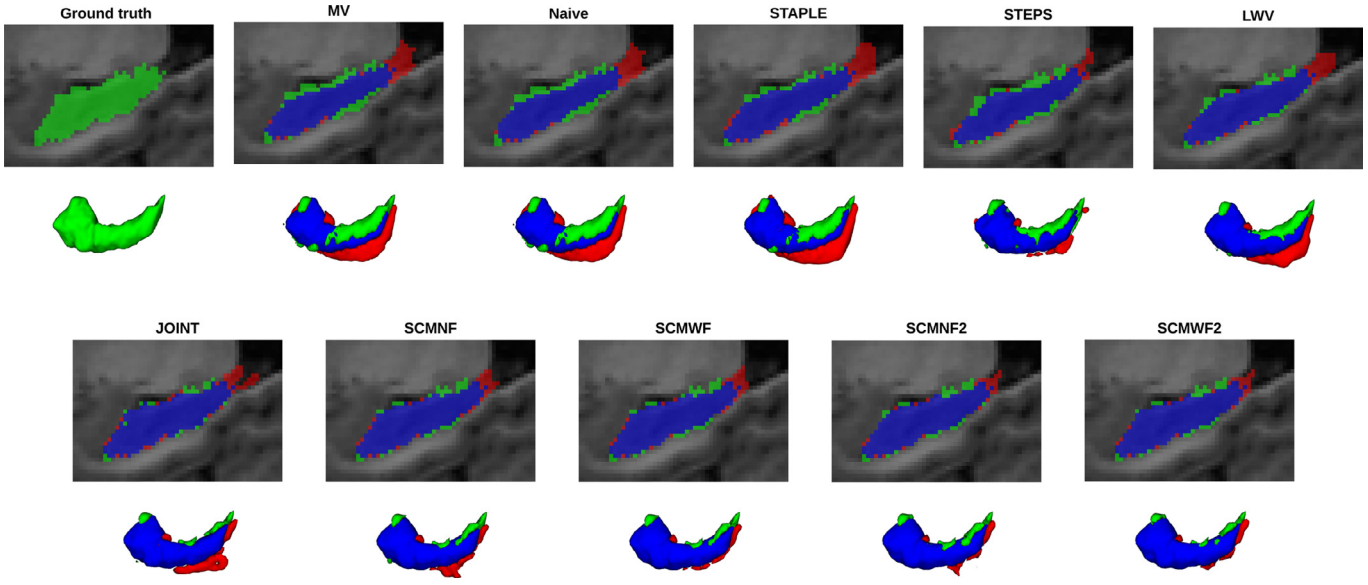

Fig. 5. Sagittal view and 3D rendering of right hippocampus segmentations for a randomly chosen image from the database. Green and red depict manual and automatic segmentations respectively. Overlap between automatic segmentation and ground truth is shown in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

SCMWF2 was statistically superior to all methods with the exception of SCMNF2 in the caudate. In the accumbens and thalamus proper, MHD provided by SCMNF2 were statistically lower than all methods except SCMWF2. Fig. 5 shows an example of automatic segmentations for all the different approaches.

Table 2 summarizes the global performance of the tested methods for all three registrations settings. In terms of overall mean Dice overlap for NR2 registration, our Naive approach (0.850) and MV (0.849) outperformed STAPLE (0.844), while it offered a slight improvement in MHD (3.299, 3.301, and 3.282 mm for MV, Naive and STAPLE). Still, with AF and NR1 registrations, STAPLE segmentations were better according to the evaluation measures. In STEPS, the benefit of using intensity information to drive label fusion was manifested in a superior performance over STAPLE for all registration settings with an 0.8%, 1.2% and 0.3% improvement in Dice overlap for AF, NR1 and NR2 registrations respectively. Furthermore, STEPS also outperformed LWV in the segmentation of the hippocampi from ADNI, although LWV segmentations in the SATA dataset were better than segmentations from STEPS for all registrations (except in MHD for the AF registration).

Regardless of the registration setting, we have to emphasize the results of four methods: JOINT, SCMWF, SCMNF2 and SCMWF2. They all produced very robust segmentations, although with some distinctions. For the NR2 registration, as illustrated in Table 1, our SCMWF2 approach yielded the highest Dice scores with 1.2%, 1.9%, 0.6% increase over JOINT in the accumbens, caudate and hippocampus (in both datasets) respectively. JOINT outperformed our approaches in the pallidum and putamen, although the improvement was minor (i.e., 0.2% in both structures). Regarding MHD, JOINT produced the lowest values only for the amygdala

(around 0.1 mm lower), whereas our approaches (i.e., SCMNF2, and SCMWF2) achieved the lowest distances in all remaining structures. The largest difference in MHD occurred in the caudate, with approximately 1 mm improvement of SCMWF2 over JOINT. Overall, SCMWF2 showed the best performance for all registration settings, as shown in Table 2. SCMWF2 was statistically superior to JOINT when using the AF registration setting in both Dice and MHD, and only in MHD when using NR2 for SATA and NR1 for ADNI. Most noteworthy is that when using coarse registrations, such as AF, Dice scores of SCMWF2 were 1% and 0.8% superior to JOINT in SATA and ADNI databases respectively. While when using finer registrations, as is the case of NR2, the improvement reduced to 0.5% and 0.6%. This demonstrates that our approach is more robust against registration failures when compared to the rest of methods. To better illustrate the robustness of our approach, Fig. 6 displays boxplots of Dice and MHD for each structure separately and each registration setting, comparing SCMWF2 and JOINT. As we can observe, the more accurate the registration, the better the segmentations, and generally, the lower the performance gap between both approaches.

5. Discussion

Our approach relies on the assumption that systematic segmentation errors caused by registration failures are atlas-dependent, and therefore can be diminished if the appearance patterns that lead to such errors are learned in each atlas space, taking into consideration the registration model. In what follows, we present a methodological comparison with the state of the art, and discuss the strengths and weaknesses of the proposed approach.

5.1. Learning from segmentation errors

There is a few number of works in the MAS literature that approach label fusion considering segmentation errors. The concepts of atlas accuracy map proposed in Sdika (2010) and reliability map proposed in Wan et al. (2008) were computed by co-registering the training atlases. However, both approaches ignored intensity information. Moreover, Wan et al. (2008) performed label fusion only for voxels where the corresponding confidence was superior to a predefined threshold, leaving unlabeled the rest of ambiguous voxels. The main drawback of these approaches is that these maps are static and may incur poor generalization if the target images are considerably different from the atlases in the training set. The risk of overffiting is also present when using offline learning (Iglesias and Sabuncu, 2015). Nonetheless, our confidence estimators do take into account the target patch appearance to compute the local confidences, and in the SCMNF2 and SCMWF2 versions we further select the most similar patch from the training atlases prior to feeding the difference between the atlas and target patches to the confidence estimator.

Supervised learning segmentation approaches existing in the literature (Hao et al., 2014; Bai et al., 2015; Sdika, 2015) learn directly from the labels, gathering the patches from the different atlases to train their classifiers. To the best of our knowledge, the method proposed by Wang et al. (2011) is the only work that shares similarities with our approach in that both methods learn from segmentation errors instead of labels. However, the wrapper method learns the disagreements between the segmentation produced by a particular host method and the ground truth segmentation in the space of the target images. Whereas in our case, we learn local confidence parameters for each atlas individually as part of a probabilistic label fusion framework.

5.2. The benefit of intensity in segmentation accuracy

As already shown in Section 4.4, incorporating intensity information in the label fusion process grants superior performance with regard to the rest of methods (e.g., MV and STAPLE). This is further emphasized by the substantial performance gain of our SCM-based approaches over the Naive method. Nevertheless, using solely intensity-based similarity, without accounting for other factors, is not enough to provide the best segmentations, as illustrated by the performance gap between LWV and, for instance, JOINT. This latter method uses a patch-based weighted voting approach where weight assignment is based on modeling dependencies between pairs of atlas patches and the target image, with the purpose of reducing the confidence of correlated erroneous atlas votes. In our approach, weight assignment accounts for the segmentation errors produced by each atlas after co-registering the remaining atlases. Besides, intensity samples used for training the confidence estimators are augmented with label-dependent features in the case of SCMWF and SCMWF2.

5.3. Similarity-based confidence estimation

Similarity-based approaches employ heuristic measures that may not be directly related to segmentation accuracy. Yet, these approaches (e.g., Coupé et al., 2011) have demonstrated excellent results in MAS. Therefore, in SCMNF2 and SCMWF2, a similarity-based approach is used in combination with supervised learning to build the confidence estimators. In SCMNF and SCMWF, an atlas patch had a single static label (i.e., the corresponding label of the central voxel from the expert segmentation). Segmentation errors were then obtained by comparing this label to the labels of the target atlases, disregarding any clue from the intensity patches. By adopting the many-to-many correspondences scheme, we equipped the atlas with information to decide what label from its surrounding neighborhood corresponds to a particular target patch. Strictly speaking, segmentation errors here are based on a specific similarity measure. Hence, what our confidence estimators try to learn is not the segmentation errors as known in SCMNF (and SCMWF), but the segmentation errors produced by an atlas through employing this specific similarity measure.

This may seem computationally more costly than the procedure used in SCMNF since we introduce an additional intermediate stage (i.e., k nearest neighbors). Nonetheless, k nearest neighbors did not suppose an important overhead and learning times were similar as mentioned in Section 4.3. Additionally, SCMNF2 demonstrated the benefits of this approach by yielding segmentation results comparable to SCMWF, with SCMWF2 outperforming SCMWF in all structures. The proposed approach uses a simple and fast classifier (i.e., logistic regression). In the state of the art, existing MAS approaches used SVM (Hao et al., 2014; Bai et al., 2015; Sdika, 2015) and random forest (Wang et al., 2014) to learn their local classifiers. For example, the learning-based approach proposed by Bai et al. (2015) used SVM with the radial basis function kernel. Thus, using a simpler model, such as logistic regression in our case, can lead to reduced training times, especially when thousands of local classifiers are to be learned.

The choice of the classifier, however, is not straightforward and is application dependent. Its performance may depend on several factors including: image modalities, nature of the features, number of samples, *etc.* Given the modularity of the proposed method, other supervised learning approaches can be used to learn our confidence estimators. Deep learning, for instance, is gaining an increasing interest in medical image analysis (Litjens et al., 2017). In our case, with deep learning, we can take advantage of the 3D nature of the image patches rather than representing the patches

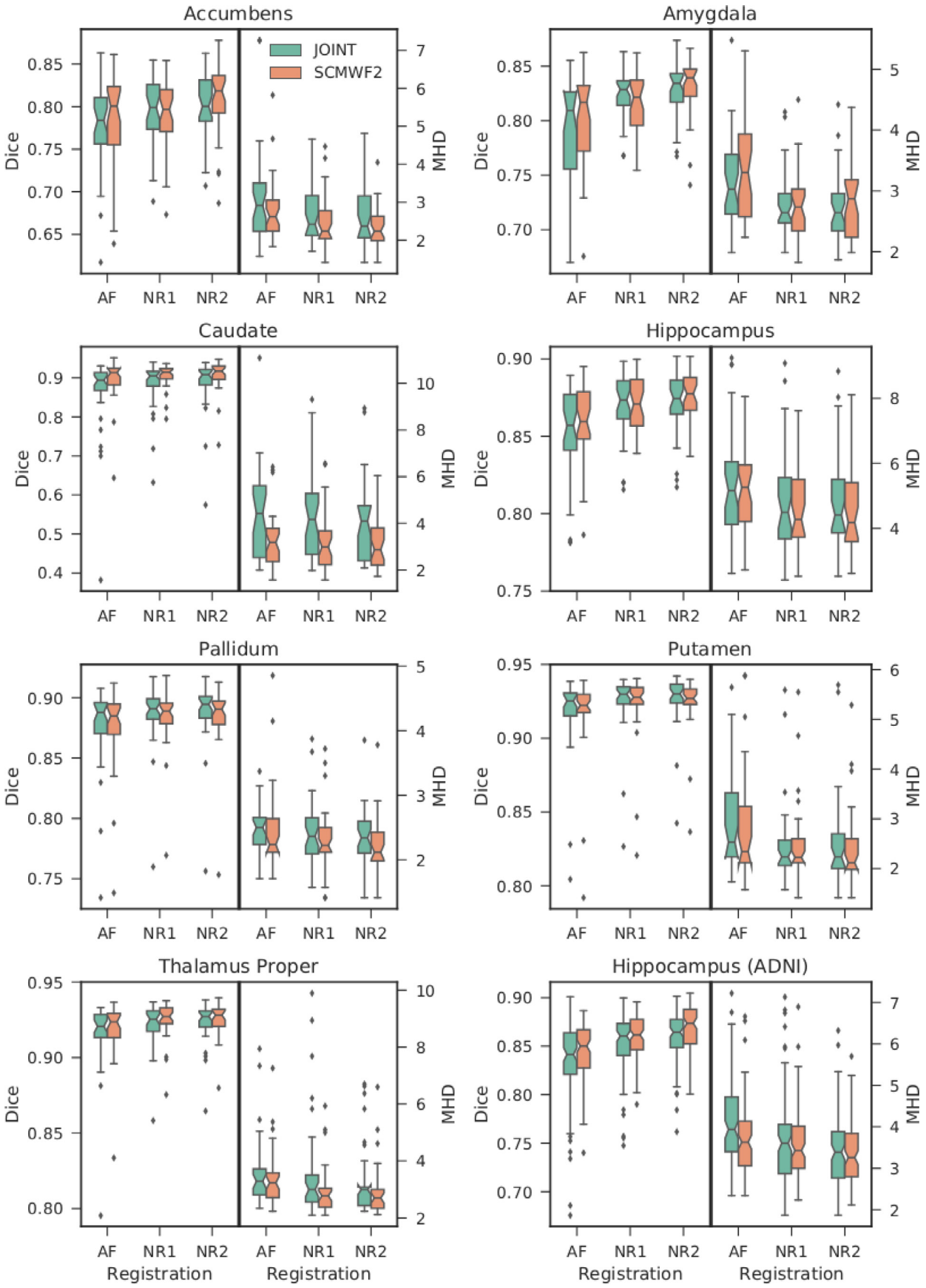

Fig. 6. Boxplots of Dice and MHD for each subcortical structure comparing SCMWF2 (orange) and JOINT (green). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

as feature vectors (Ciresan et al., 2012; de Brébisson and Montana, 2015). Beyond patch-based approaches, architectures such as the U-net proposed by Ronneberger et al. (2015), and its 3D extensions (Çiçek et al., 2016), can also be used as global confidence estimators since they can take whole images as input and output a classification for each pixel/voxel, which permits to take into consideration larger contextual information for each voxel by analyzing the images at multiple scales. Although more time-consuming than logistic regression, this is a promising line of future work.

5.4. The effect of label-dependent features

By incorporating label-dependent features, results showed that both SCMWF and SCMWF2 can output segmentations similar to state-of-the-art methods for all subcortical brain structures used in our experiments. The benefits of using additional features beyond patch intensity is well demonstrated in Asman et al. (2015), Bai et al. (2015), Hao et al. (2014) and Wang et al. (2011). In Asman et al. (2015), 1009 dimensional feature vectors were used. Similarly, high-dimensional vectors of 1003 features were used by Wang et al. (2011). For voxel-wise classifiers, Hao et al. (2014) used feature vectors of 379 elements and SVM with l1-regularization to select the sparsest solution from all the possibly redundant features. In Bai et al. (2015), 260-element samples were built from intensity, gradient and contextual features. In our case, SCMNF and SCMNF2 used feature vectors of only 27 dimensions (i.e., intensity patch of $3 \times 3 \times 3$), whereas augmented feature vectors of 33 elements (i.e., 6 additional label-dependent features) were used in SCMWF and SCMWF2. The informative power of the label-dependent features can be observed in the considerable increase in performance of SCMWF with respect to SCMNF. However, the boost achieved by SCMWF2 with regard to SCMNF2 is not that large, except in the segmentation of the hippocampus from the ADNI dataset, with 1.3% increase in Dice and a 0.14 mm decrease in MHD approximately. This is possibly due to the fact that we are reaching inter-rater variability is some structures (e.g., thalamus proper). Moreover, more compact segmentations were obtained by SCMNF2 than SCMWF2 for the accumbens, pallidum, putamen and thalamus proper in terms of MHD for the NR2 registration, as shown in Table 1.

The idea of using information from the label patch was already used in Wang et al. (2011) by appending to the feature vectors the segmentation labels produced by the host method in the neighborhood of each voxel. However, they just used the raw label patch as features. Our label-dependent feature extraction procedure is different from all the aforementioned approaches. Instead of predefined filters, we use the atlas label patch as a mask to compute the difference between the features extracted from each region of the intensity patch. This reduced number of features seemed to provide the local classifiers with potential information to better discriminate between correct and erroneous atlas patches.

Feature extraction is computationally expensive, especially when using the non-local approach with voxel-wise classifiers. As reported in Section 4.3, offline learning took 3 min for SCMNF, with a runtime increase of 7 min for SCMWF when incorporating the label-dependent features. Considering sample selection strategies may turn out advantageous to decrease computational cost. Bai et al. (2015), for instance, performed patch selection to reduce such computations.

5.5. The influence of outliers

Target images that highly deviate from the anatomies in the training set have a negative impact in registration, giving rise to misaligned structures (i.e., outliers). However, our approach has proven its robustness against outliers since SCMWF2 achieved the best segmentations with the lowest Dice and MHD standard deviations for all registration settings and both SATA and ADNI datasets, as shown in Table 2. Moreover, the similarity-based label fusion approaches used in the conducted experiments (i.e., LWV and JOINT) seem to perform poorly in the segmentation of outliers, as illustrated in Fig. 7. This figure shows the structure with the worst segmentations provided by all methods. The Dice scores achieved by LWV, JOINT, SCMNF2 and SCMWF2 in segmenting this structure using AF registration are 0.150, 0.217, 0.396 and 0.417 (0.415, 0.486, 0.692 and 0.701 for NR1 registration), respectively. Note that overlaps provided by all methods when using the affine registration are below 0.5, and for NR1 registration both LWV and JOINT Dice overlaps are below 0.5. We can observe that this is a clear failure in registration caused by the enlarged ventricles next to the caudate. Segmentation errors produced by the aforementioned similarity-based approaches are mainly due to over-segmentation (i.e., red color). Especially important is the over-segmentation produced by identifying part of the left ventricle as caudate, although there is great difference in intensity because the ventricles have a lower intensity range. This is further illustrated by the Dice and MHD boxplots of the caudate in Fig. 6. Note, however, that results shown in the boxplots do not coincide with the Dice scores of this outlier because the boxplots were created by averaging the results achieved for the left and right caudate. Our confidence estimators seem to show more robustness in the presence of outlier patches. Another clear impact of the presence of outliers in performance of JOINT can be seen in the boxplots of the thalamus proper and the hippocampus from ADNI, notably with the affine registration.

5.6. Robustness to registration failures

The claim of our work is that systematic segmentation errors due to registration can be substantially mitigated using the proposed approach. Thus, we studied the effect of registration in segmentation results in order to asses how segmentation performance evolves from using more coarse (i.e., AF) to finer registrations (i.e., NR2). From the overall segmentation performance reported in Table 2 and boxplots in Fig. 6, we can conclude that our method is more robust to registration errors. With coarse registrations, which are more prone to failures, our approach achieved the largest performance increment compared to the rest of approaches. Therefore, demonstrating to be robust to registration failures. Overall mean Dice scores obtained by SCMWF2 with AF registration were lower (0.865 overlap in SATA and 0.843 in ADNI) than the ones obtained with NR2 (0.880 in SATA and 0.866 in ADNI). The same occurs with MHD, with SCMWF2 providing larger distances (3.312 mm in SATA and 3.728 in ADNI) when using AF than the distances achieved with NR2 (2.898 mm in SATA and 3.369 in ADNI). This difference in the performance of our approach between AF and NR2, could be substantially reduced by taking advantage of the many-to-many correspondences scheme and using larger patch and window search sizes, rather than the $3 \times 3 \times 3$ size used in this work.

5.7. Limitations and future directions

The main limitation of our approach is that confidence learning is performed for each training atlas, which makes it computationally expensive. One possible way to lessen this computational burden is to restrict the learning process to the most representative atlas spaces, for example, by clustering the atlases and only learning in the centroid spaces. To further reduce computational time, although at the expense of sacrificing segmentation accuracy, clustering can also be applied to learn considering groups of neighboring voxels instead of using voxel-wise classifiers. On the other hand, in this work, the size of intensity patches and search neighborhoods used in both target an atlas spaces was set to $3 \times 3 \times 3$

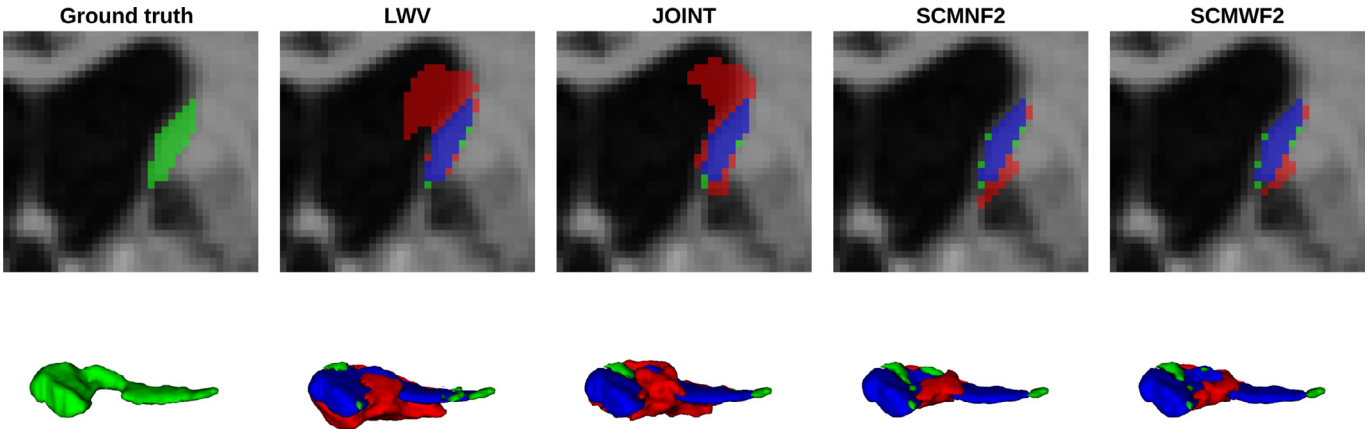

Fig. 7. Illustration of left caudate automatic segmentation with the lowest Dice score and largest MHD using NR1 registration. Green and red depict manual and automatic segmentations respectively. Overlap is depicted in blue. The first row shows ground truth and automatic segmentations in coronal view. The second row shows the corresponding 3D renderings. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(i.e., radius of 1). As future work, the impact of this parameter in segmentation accuracy can be studied. Likewise, parameter tuning can be performed to find the best value for the penalty parameter used in logistic regression. This parameter can be optimized in a local manner for each of the voxel-wise classifiers or for the whole structure.

The methods based on the many-to-many correspondences scheme yielded very accurate results. Still, more work is required here to asses the effect of the similarity metric and the strategy used to predict the label, considering only the most similar patch or some heuristic weighting the patches contribution to the final label. Another direction of future work is to explore different strategies to extract our label-dependent features or even adopt a supervised approach to learn such features rather than using feature engineering. Moreover, additional features as the ones used in Hao et al. (2014) and Bai et al. (2015) could be considered. Finally, a very promising direction of future work is to consider correlations between voxels and/or the votes of the training atlases (Wang et al., 2014).

6. Conclusions

Registration failures constitute a potential source of systematic errors in MAS. In this manuscript, we have proposed a probabilistic label fusion framework that takes into consideration local atlas confidences at each point by the estimation of the so-called spatial confidence maps. Given the nature of our approach, we have also proposed a novel label-dependent feature extraction that provided valuable information in the prediction of the confidences. Systematic errors due to registration are accounted for during label fusion since confidence learning is performed in atlas space. As opposed to STAPLE-like approaches, this learning process is performed in an offline manner using the available training atlases. Therefore, computational complexity at test time is comparable to the simplest approaches. Furthermore, incorporating neighborhood information in atlas space to compute the segmentation errors rendered our approach more robust to registration errors. Experimental results have shown that our approach yields superior performance to state-of-the-art approaches in the segmentation of the majority of subcortical brain structures.

Acknowledgments

This work is co-financed by the Marie Curie FP7-PEOPLE-2012-COFUND Action, Grant agreement no: 600387.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.media.2017.08.008

References

Aljabar, P., Heckemann, R., Hammers, A., Hajnal, J., Ruecker, D., 2009. Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. NeuroImage 46 (3), 726–738. doi:10.1016/j.neuroimage.2009.02.018.

Artaechevarria, X., Muñoz Barrutia, A., Ortiz-de Solórzano, C., 2009. Combination strategies in multi-atlas image segmentation: Application to brain MR data. IEEE Trans. Med. Imaging 28 (8), 1266–1277.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. NeuroImage 26 (3), 839–851. doi:10.1016/j.neuroimage.2005.02.018.

Asman, A.J., Huo, Y., Plassard, A.J., Landman, B.A., 2015. Multi-atlas learner fusion: An efficient segmentation approach for large-scale data. Med. Image Anal. 26 (1), 82–91. doi:10.1016/j.media.2015.08.010.

Asman, A.J., Landman, B.A., 2012. Formulating spatially varying performance in the statistical fusion framework. IEEE Trans. Med. Imaging 31 (6), 1326–1336. doi:10.1109/TMI.2012.2190992.

Asman, A.J., Landman, B.A., 2013. Non-local statistical label fusion for multi-atlas segmentation. Med. Image Anal. 17 (2), 194–208. doi:10.1016/j.media.2012.10.002.

Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. 12 (1), 26–41. doi:10.1016/j.media.2007.06.004.

Bai, W., Shi, W., Ledig, C., Rueckert, D., 2015. Multi-atlas segmentation with augmented features for cardiac {MR} images. Med. Image Anal. 19 (1), 98–109. doi:10.1016/j.media.2014.09.005.

Benkarim, O.M., Piella, G., González Ballester, M.A., Sanroma, G., 2016. Enhanced Probabilistic Label Fusion by Estimating Label Confidences Through Discriminative Learning. Springer International Publishing, Cham, pp. 505–512. doi:10.1007/978-3-319-46723-8_58.

Benkarim, O.M., Radeva, P., Igual, L., 2014. Label Consistent Multiclass Discriminative Dictionary Learning for MRI Segmentation. Springer International Publishing, Cham, pp. 138–147. doi:10.1007/978-3-319-08849-5_14.

de Brébisson, A., Montana, G., 2015. Deep neural networks for anatomical brain segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 20–28. doi:10.1109/CVPRW.2015.7301312.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Buades, A., Coll, B., Morel, J.M., 2005. A non-local algorithm for image denoising. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2, pp. 60–65. doi:10.1109/CVPR.2005.38.

Cardoso, J.M., Leung, K., Modat, M., Keihaninejad, S., Cash, D., Barnes, J., Fox, N.C., Ourselin, S., 2013. Steps: Similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation. Med. Image Anal. 17 (6), 671–684.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. Springer International Publishing, Cham, pp. 424–432. doi:10.1007/978-3-319-46723-8_49.

Ciresan, D.C., Gambardella, L.M., Giusti, A., Schmidhuber, J., 2012. Deep neural networks segment neuronal membranes in electron microscopy images. In: In NIPS, pp. 2852–2860.

Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L., 2011. Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. NeuroImage 54 (2), 940–954.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302. doi:10.2307/1932409.

Dubuisson, M.P., Jain, A.K., 1994. A modified hausdorff distance for object matching. In: Proceedings of 12th International Conference on Pattern Recognition, 1, pp. 566–568. doi:10.1109/ICPR.1994.576361.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J., 2008. Liblinear: A library for large linear classification. J. Mach. Learn. Res. 9, 1871–1874.

González-Villà, S., Oliver, A., Valverde, S., Wang, L., Zwiggelaar, R., Lladó, X., 2016. A review on brain structures segmentation in magnetic resonance imaging. Artif. Intell. Med. 73, 45–69. doi:10.1016/j.artmed.2016.09.001.

Gorthi, S., Akhondi-Asl, A., Thiran, J.-P., Warfield, S.K., 2014. Optimal MAP Parameters Estimation in STAPLE - Learning from Performance Parameters versus Image Similarity Information. Springer International Publishing, Cham, pp. 174–181. doi:10.1007/978-3-319-10581-9_22.

Hao, Y., Wang, T., Zhang, X., Duan, Y., Yu, C., Jiang, T., Fan, Y., for the Alzheimer's Disease Neuroimaging Initiative, 2014. Local label learning (LLL) for subcortical structure segmentation: Application to hippocampus segmentation. Hum. Brain Mapp. 35 (6), 2674–2697. doi:10.1002/hbm.22359.

Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A., 2006. Automatic anatomical brain {MRI} segmentation combining label propagation and decision fusion. NeuroImage 33 (1), 115–126. doi:10.1016/j.neuroimage.2006.05.061.

Iglesias, J.E., Sabuncu, M.R., 2015. Multi-atlas segmentation of biomedical images: A survey. Med. Image Anal. 24 (1), 205–219. doi:10.1016/j.media.2015.06.012.

Isgum, I., Staring, M., Rutten, A., Prokop, M., Viergever, M.A., van Ginneken, B., 2009. Multi-atlas-based segmentation with local decision fusion: Application to cardiac and aortic segmentation in ct scans. IEEE Trans. Med. Imaging 28 (7), 1000–1010. doi:10.1109/TMI.2008.2011480.

Klein, A., Mensh, B., Ghosh, S., Tourville, J., Hirsch, J., 2005. Mindboggle: automated brain labeling with multiple atlases. BMC Med. Imaging 5 (1), 7. doi:10.1186/1471-2342-5-7.

Leemput, K.V., Maes, F., Vandermeulen, D., Suetens, P., 1999. Automated model-based tissue classification of mr images of the brain. IEEE Trans. Med. Imaging 18 (10), 897–908. doi:10.1109/42.811270.

Litjens, G.J.S., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. CoRR abs/1702.05747.

Lötjönen, J.M., Wolz, R., Koikkalainen, J.R., Thurfjell, L., Waldemar, G., Soininen, H., Rueckert, D., 2010. Fast and robust multi-atlas segmentation of brain magnetic resonance images. NeuroImage 49 (3), 2352–2365. doi:10.1016/j.neuroimage.2009.10.026.

Morra, J.H., Tu, Z., Apostolova, L.G., Green, A.E., Toga, A.W., 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 (1), 30–43. doi:10.1109/TMI.2009.2021941.

Nyul, L.G., Udupa, J.K., Zhang, X., 2000. New variants of a method of mri scale standardization. IEEE Trans. Med. Imaging 19 (2), 143–150. doi:10.1109/42.836373.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Powell, S., Magnotta, V.A., Johnson, H., Jammalamadaka, V.K., Pierson, R., Andreasen, N.C., 2008. Registration and machine learning-based automated segmentation of subcortical and cerebellar brain structures. NeuroImage 39 (1), 238–247. doi:10.1016/j.neuroimage.2007.05.063.

Rohlfing, T., Brandt, R., Menzel, R., Maurer, C.R., 2004. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. NeuroImage 23 (8), 983–994.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. CoRR abs/1505.04597.

Rousseau, F., Habas, P.A., Studholme, C., 2011. A supervised patch-based approach for human brain labeling. IEEE Trans. Med. Imaging 30 (10), 1852–1862. doi:10.1109/TMI.2011.2156806.

Sabuncu, M.R., Yeo, B.T.T., Leemput, K.V., Fischl, B., Golland, P., 2010. A generative model for image segmentation based on label fusion. IEEE Trans. Med. Imaging 29 (10), 1714–1729. doi:10.1109/TMI.2010.2050897.

Sanroma, G., Benkarim, O.M., Piella, G., González Ballester, M.A., 2016. Building an Ensemble of Complementary Segmentation Methods by Exploiting Probabilistic Estimates. Springer International Publishing, Cham, pp. 27–35. doi:10.1007/978-3-319-47157-0_4.

Sanroma, G., Wu, G., Gao, Y., Shen, D., 2014. Learning to rank atlases for multiple-atlas segmentation. IEEE Trans. Med. Imaging 33 (10), 1939–1953. doi:10.1109/TMI.2014.2327516.

Sanroma, G., Wu, G., Gao, Y., Thung, K.-H., Guo, Y., Shen, D., 2015. A transversal approach for patch-based label fusion via matrix completion. Med. Image Anal. 24 (1), 135–148. doi:10.1016/j.media.2015.06.002.

Sanroma, G., Wu, G., Kim, M., González Ballester, M.A., Shen, D., 2016. Chapter 11 - Multiple-Atlas Segmentation in Medical Imaging. In: Zhou, S.K. (Ed.), Medical Image Recognition, Segmentation and Parsing. Academic Press, pp. 231–257. doi:10.1016/B978-0-12-802581-9.00011-1.

Sdika, M., 2010. Combining atlas based segmentation and intensity classification with nearest neighbor transform and accuracy weighted vote. Med. Image Anal. 14 (2), 219–226. doi:10.1016/j.media.2009.12.004.

Sdika, M., 2015. Enhancing atlas based segmentation with multiclass linear classifiers. Med. Phys. 42 (12), 7169–7181. doi:10.1118/1.4935946.

Shattuck, D.W., Sandor-Leahy, S.R., Schaper, K.A., Rottenberg, D.A., Leahy, R.M., 2001. Magnetic resonance image tissue classification using a partial volume model. NeuroImage 13 (5), 856–876. doi:10.1006/nimg.2000.0730.

Wan, J., Carass, A., Resnick, S.M., Prince, J.L., 2008. Automated reliable labeling of the cortical surface. In: Proceedings. IEEE International Symposium on Biomedical Imaging, 2008, p. 440.

Wang, H., Cao, Y., Syeda-Mahmood, T., 2014. Multi-atlas Segmentation with Learning-Based Label Fusion. Springer International Publishing, Cham, pp. 256–263. doi:10.1007/978-3-319-10581-9_32.

Wang, H., Das, S.R., Suh, J.W., Altinay, M., Pluta, J., Craige, C., Avants, B., Yushkevich, P.A., 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 (3), 968–985. doi:10.1016/j.neuroimage.2011.01.006.

Wang, H., Suh, J.W., Das, S.R., Pluta, J.B., Craige, C., Yushkevich, P.A., 2013. Multi-atlas segmentation with joint label fusion. IEEE Trans. Pattern Anal. Mach. Intell. 35 (3), 611–623.

Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (staple): An algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging 23 (7), 903–921.

Zhang, D., Guo, Q., Wu, G., Shen, D., 2012. Sparse Patch-Based Label Fusion for Multi-Atlas Segmentation. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 94–102. doi:10.1007/978-3-642-33530-3_8.

Zikic, D., Glocker, B., Criminisi, A., 2013. Atlas encoding by randomized forests for efficient label propagation. In: MICCAI 2013 - 16th Intl Conf. on Medical Image Computing and Computer Assisted Intervention. Springer.