NeuroImage 106 (2015) 34–46

Contents lists available at ScienceDirect

NeuroImage

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

Hierarchical multi-atlas label fusion with multi-scale feature representation and label-specific patch partition ☆

Guorong Wu [a], Minjeong Kim [a], Gerard Sanroma [a], Qian Wang [b], Brent C. Munsell [c], Dinggang Shen [a,d,*], The Alzheimer's Disease Neuroimaging Initiative

[a] BRIC and Department of Radiology, University of NC, Chapel Hill, USA
[b] Med-X Research Institute, Shanghai Jiao Tong University, Shanghai, China
[c] Computer Science Department, College of Charleston, Charleston, SC 29424, USA
[d] Department of Brain and Cognitive Engineering, Korea University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 6 May 2014
Revised 5 November 2014
Accepted 12 November 2014
Available online 20 November 2014

Keywords:
Patch-based labeling
Multi-atlas based segmentation
Multi-scale feature representation
Label-specific patch partition
Sparse representation

ABSTRACT

Multi-atlas patch-based label fusion methods have been successfully used to improve segmentation accuracy in many important medical image analysis applications. In general, to achieve label fusion a single target image is first registered to several atlas images. After registration a label is assigned to each target point in the target image by determining the similarity between the underlying target image patch (centered at the target point) and the aligned image patch in each atlas image. To achieve the highest level of accuracy during the label fusion process it's critical for the chosen patch similarity measurement to accurately capture the tissue/shape appearance of the anatomical structure. One major limitation of existing state-of-the-art label fusion methods is that they often apply a fixed size image patch throughout the entire label fusion procedure. Doing so may severely affect the fidelity of the patch similarity measurement, which in turn may not adequately capture complex tissue appearance patterns expressed by the anatomical structure. To address this limitation, we advance state-of-the-art by adding three new label fusion contributions: First, each image patch is now characterized by a multi-scale feature representation that encodes both local and semi-local image information. Doing so will increase the accuracy of the patch-based similarity measurement. Second, to limit the possibility of the patch-based similarity measurement being wrongly guided by the presence of multiple anatomical structures in the same image patch, each atlas image patch is further partitioned into a set of label-specific partial image patches according to the existing labels. Since image information has now been semantically divided into different patterns, these new label-specific atlas patches make the label fusion process more specific and flexible. Lastly, in order to correct target points that are mislabeled during label fusion, a hierarchical approach is used to improve the label fusion results. In particular, a coarse-to-fine iterative label fusion approach is used that gradually reduces the patch size. To evaluate the accuracy of our label fusion approach, the proposed method was used to segment the hippocampus in the ADNI dataset and 7.0 T MR images, sub-cortical regions in LONI LBPA40 dataset, mid-brain regions in SATA dataset from MICCAI 2013 segmentation challenge, and a set of key internal gray matter structures in IXI dataset. In all experiments, the segmentation results of the proposed hierarchical label fusion method with multi-scale feature representations and label-specific atlas patches are more accurate than several well-known state-of-the-art label fusion methods.

Introduction

Many medical image analysis studies require an accurate segmentation of anatomical structures in order to measure structural differences across individuals or between groups (Aljabar et al., 2009; Hsu et al., 2002). For example, in connectome applications multiple brain regions, in hundreds of brain MR images, need to be automatically identified before constructing a brain connectivity network (Liu et al., 2012; Liu and Ye, 2010) that describes network architecture of the human brain. Therefore, to improve segmentation accuracy the development of automatic ROI (region of interest) labeling methods has seen increased

☆ Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu/). As such, the investigators with 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.

* Corresponding author at: BRIC and Department of Radiology, University of North Carolina, Chapel Hill, USA. Fax: +1 919 843 2641.
E-mail address: dgshen@med.unc.edu (D. Shen).

http://dx.doi.org/10.1016/j.neuroimage.2014.11.025
1053-8119/

attention in the medical imaging field over the last several years (Aljabar et al., 2009; Coupé et al., 2011; Rousseau et al., 2011; Tong et al., 2012; Wang et al., 2011a; Wang et al., 2011b; Warfield et al., 2004; Wu et al., 2014).

Multiple atlases with manually identified labels have proven to be very useful when used to detect and label ROIs in the target image that may show high structural variations in the population. The basic assumption behind multi-atlas based segmentation is that the target image point should bear the same label as the atlas image point if the local tissue shape or appearance is very similar. All atlas images are required to be registered to a target image before label fusion. To alleviate possible registration errors, patch-based label fusion (Coupé et al., 2011; Rousseau et al., 2011) seeks multiple correspondence candidates using patchwise similarity measurements between the target image patch and the atlas image patches within a certain voxel neighborhood. Intuitively, if the calculated similarity measurement between a target image patch and a particular atlas image patch is very high, then the atlas label assigned to the target point is the correct one.

To accurately assess image patch similarity, the identification and selection of ideal image patches are key components of patch-based label fusion methods. Most state-of-the-art methods simply use fixed size patches throughout the entire label fusion procedure. For example, $7 \times 7 \times 7$ or $9 \times 9 \times 9$ cubic patches are usually used in the literature (Coupé et al., 2011; Rousseau et al., 2011; Tong et al., 2012; Wang et al., 2011a). In order to make the label fusion robust to noise, image patches are required to be sufficiently large enough to capture the intended image content. However, using a large image patch may create additional problems when labeling small anatomical structures, e.g. the patchwise similarity measurement could be dominated by other larger anatomical structures surrounding the smaller one in the image patch. In short, methods that use fixed-size patches lack discriminative power to characterize complex appearance patterns in the medical imaging data.

During the last decade, many efforts have been made to improve the discrimination ability of image patches during label fusion. For instance, sparse dictionary learning is used in Tong et al. (2013) to find the best feature representations prior to label fusion. Moreover, in Wang et al. (2011a) and Wu et al. (2014) dependencies among atlas image patches have been investigated to improve labeling accuracy by iteratively inspecting incorrectly labeled patches that show similar labeling error patterns. However, these state-of-the-art approaches use patches with fixed size and therefore still suffer from this limitation.

In this paper, we address the above limitations by developing hierarchical and high-level feature representations to adequately describe image patches. We propose the following three contributions: *First*, a layer-wise multi-scale feature representation adaptively encodes image features at different scales for each image point in the image patch. In the proposed approach, feature representations near the center of the patch provide more detailed (fine-scale) shape or appearance information, whereas feature representations near the edge of the patch provide less detailed (coarse-scale) shape or appearance information. *Second*, it's very common that the structure to be segmented, e.g. the hippocampus, is surrounded by other anatomical structures in the image patch. In such cases it becomes very difficult to correctly recognize the intended structure from the surrounding ones and mislabeling is likely to occur. In computer vision, object recognition algorithms address this limitation by attempting to separate the foreground pattern from background clutter (Li et al., 2010). In light of this research, a novel label-specific patch partition technique is proposed that splits each atlas patch into a set of new complementary label-specific (or structure-specific) image patches. To handle the increased number of label-specific image patches after the proposed patch splitting strategy a group sparsity constraint is included. As a result, the discriminative power of each label-specific image patch is enhanced because it only contains the image information of the corresponding anatomical structure. To the best of our knowledge, this type of representation is rarely exploited in label fusion. *Third*, because existing label fusion methods typically use a fixed patch size, and label the entire target image in one pass, they are not given a chance to correct possible errors. To overcome this limitation the proposed method uses an iterative label-fusion procedure. Specifically, larger image patches are used in the beginning to increase the search range, however at each iteration the labeling result is evaluated and the size of the image patch is gradually reduced. To ensure that spurious artifacts do not dominate the proposed label-fusion method, a sparsity constraint is included that only allows a small number of atlas patches to participate in the label fusion process.

It should be noted that this paper is an extension of our previous work in Wu and Shen (2014). However, there are several differences, specifically: a group sparsity constraint is used instead of a weighting vector sparsity constraint, a more comprehensive validation of each contribution (i.e., multi-scale feature representation, label-specific patch partition, and iterative label fusion), and additional datasets are used to evaluate the performance of the proposed label fusion method.

Performance of the proposed label fusion method is compared to existing state-of-the-art patch-based labeling methods (Coupé et al., 2011; Rousseau et al., 2011) using several different datasets. Specifically, the datasets used to evaluate the proposed method are the MICCAI 2013 segmentation challenge dataset (Landman and Warfield, 2012) with 14 manually labeled ROIs in the mid-brain, the LONI LBPA40 dataset (Shattuck et al., 2008) with 54 manually labeled ROIs at subcortical regions, and the IXI dataset with 83 manually labeled ROIs (Hammers et al., 2003; Hammers et al., 2007). Finally, we also include hippocampus segmentation experiments using the ADNI (Alzheimer's Disease Neuroimaging Initiative) dataset and 7.0 T MR images (Cho et al., 2010). For each dataset the proposed method achieves a more accurate labeling result.

The remainder of the paper is organized as follows: In the Method section we present our novel generative probability model for label fusion, in the Experiments section we evaluate its performance by comparing it with conventional patch-based methods, and in the Discussion section we provide a brief conclusion.

Method

Given the target image $\boldsymbol{T}$, the goal of label fusion is to automatically determine the label map $\boldsymbol{L_T}$ for the target image. We first register each atlas image, as well as the label maps, onto the target image space. We use $\boldsymbol{I} = \{\boldsymbol{I_s} | \boldsymbol{s} = 1, \ldots, \boldsymbol{N}\}$ and $\boldsymbol{L} = \{\boldsymbol{L_s} | \boldsymbol{s} = 1, \ldots, \boldsymbol{N}\}$ to denote the $\boldsymbol{N}$ registered atlases and label maps, respectively. For each target image point $\boldsymbol{x}$ ($\boldsymbol{x} \in \boldsymbol{T}$), all the atlas patches* within a certain search neighborhood $\boldsymbol{n}(\boldsymbol{x})$, denoted as $\vec{\boldsymbol{\beta}}_{\boldsymbol{s},\boldsymbol{y}}$ ($\vec{\boldsymbol{\beta}}_{\boldsymbol{s},\boldsymbol{y}} \subset \boldsymbol{I_s}, \boldsymbol{y} \in \boldsymbol{n}(\boldsymbol{x})$), are used to compute the patchwise similarities w.r.t. the target image patch $\vec{\boldsymbol{\alpha}}_{\boldsymbol{T},\boldsymbol{x}}$ ($\vec{\boldsymbol{\alpha}}_{\boldsymbol{T},\boldsymbol{x}} \subset \boldsymbol{T}$). We arrange each patch, $\vec{\boldsymbol{\beta}}_{\boldsymbol{s},\boldsymbol{y}}$ and $\vec{\boldsymbol{\alpha}}_{\boldsymbol{T},\boldsymbol{x}}$, into a column vector. We use the tuple $\boldsymbol{b} = (\boldsymbol{s}, \boldsymbol{y})$ to denote both the atlas image index $\boldsymbol{s}$ and the location of the patch center point $\boldsymbol{y}$, respectively. Thus, each atlas image patch $\vec{\boldsymbol{\beta}}_{\boldsymbol{s},\boldsymbol{y}}$ can now be simplified to $\vec{\boldsymbol{\beta}}_{\boldsymbol{b}}$ ($\boldsymbol{b} = 1, \ldots, \boldsymbol{Q}$), where $\boldsymbol{Q} = \boldsymbol{N} \times |\boldsymbol{n}(\boldsymbol{x})|$ is the total number of atlas image patches which are used to label the center point of the target image patch $\vec{\boldsymbol{\alpha}}_{\boldsymbol{T},\boldsymbol{x}}$. For clarity, we use only $\vec{\boldsymbol{\alpha}}$ to denote the underlying target image patch by dropping off the subscripts in $\vec{\boldsymbol{\alpha}}_{\boldsymbol{T},\boldsymbol{x}}$.

Label fusion methods such as non-local averaging (Coupé et al., 2011; Rousseau et al., 2011), can be used to calculate the weighting vector $\vec{w} = [w_b]_{b=1,\ldots,Q}$ for all atlas patches, each of which is denoted by $\vec{\beta}_b$. As we will explain in the Label-specific Atlas Patch Partition section, we adopt the sparsity constraint (Liu et al., 2009a,b; Tibshirani, 1996) in our method by regarding the label fusion procedure as the problem of

* Some label fusion methods use patch pre-selection to discard the less similar patches.

finding the optimal combination among a set of atlas patches $\{\vec{\beta}_b\}$ for the target image patch $\vec{\alpha}$ (Tong et al., 2012; Zhang et al., 2012):

$$\hat{\vec{w}} = \arg\min_{\vec{w}} \frac{1}{2} \| \vec{\alpha} - \boldsymbol{B}\vec{w} \|^2 + \lambda \| \vec{w} \|_1, \quad (1)$$

where the scalar $\boldsymbol{\lambda}$ controls the strength of sparsity constraint and $\boldsymbol{B}$ is a matrix built by assembling all column vectors $\{\vec{\boldsymbol{\beta}}_{\boldsymbol{b}}\}$ in a columnwise manner. The image patch vectors are usually required to be normalized to the unit vector before optimizing over the sparse coefficients $\vec{\boldsymbol{w}}$ (Wright et al., 2009). Assuming that we have $\boldsymbol{M}$ possible labels $\{\boldsymbol{l}_1, \ldots, \boldsymbol{l}_{\boldsymbol{m}}, \ldots, \boldsymbol{l}_{\boldsymbol{M}}\}$ in the atlases, the label on target image point $\boldsymbol{x}$ can be efficiently determined by:

$$\hat{L}_T(x) = \arg\max_{m=1,\ldots,M} \sum_{b=1}^{Q} [w_b \cdot \delta(L_b, l_m)], \quad (2)$$

where L_b denotes the label in the center point of the atlas patch β_b, and the Dirac function $\delta(L_b, l_m)$ is equal to 1 when $L_b = l_m$ and 0 otherwise.

As we can see in Eq. (1), the image intensities in the entire image patch are used for label fusion. Since one image patch may contain more than one anatomical structure and the to-be-segmented target ROI may have a complex shape/appearance pattern, the current patch-based label fusion methods have a certain risk of being misled by the patchwise similarities computed using image patches of fixed size or scale. We address this issue by introducing the idea of adaptive scale that has the following three components. *Firstly*, we treat each-element within the image patch differently w.r.t. the radial distance toward the patch center. Therefore, a single image patch can convey image information from multiple scales (section Multi-scale feature representations). *Secondly*, we treat the label information within the image patch separately, instead of as a whole. Specifically, we adaptively build label-specific atlas patches by using the existing label information in the atlases (section Label-specific atlas patch partition). *Thirdly*, we dynamically reduce the patch size from large to small in order to hierarchically improve the label fusion accuracy in a coarse-to-fine manner (section Hierarchical patch-based label fusion).

Multi-scale feature representations

As demonstrated in our previous work (Wu et al., 2006), image points at different brain regions should use different image scales to precisely characterize the local anatomical information. However, in most patch-based label fusion methods, every point in the image patch contributes equally and uses just its own intensity value for computing the patchwise similarity. We overcome this limitation by allowing each point to use an adaptive scale for capturing local appearance characteristics. Specifically, we first partition the whole image patch into several nested non-overlapping *layers*, spreading from the center point to the boundaries of the image patch. Next, we capture the fine-scale features for the layer closest to the patch center since the label fusion procedure eventually aims at determining the label for the central point. We gradually use larger and larger scales to capture the coarse-scale information as the distance to the patch center increases. Although the image pyramid technique (Liu and Ye, 2010) can be applied for multi-scale feature representation, we choose the less computationally demanding solution of adaptively replacing the original intensity values with the convolved intensity values using different Gaussian filters.

Fig. 1 illustrates the procedure of how to integrate the multi-scale feature representation into the conventional image patch. In the following example, we use three non-overlapping layers. First, we deploy three Gaussian filters upon the original image patch separately and obtain three smoothed image patches. Then, for each element in the image patch, we replace its original intensity value with the new value in the smoothed image at the same location. In this example, we replace the intensities in the inner most layer by the convolved intensity values smoothed via a Gaussian filter with the smallest kernel (in blue, in the right side of the figure). For each point in the middle layer, we use the convolved intensity value from the smoothed image patch via a Gaussian filter with the medium kernel (in red). Similarly, we use the smoothed image patch via a Gaussian filter with the largest kernel as the feature representation for the image points in the third layer (in green). In this way, the image patch is now equipped with the multi-scale feature representations, as shown in the right of Fig. 1. Hereafter, $\vec{\alpha}$ and $\vec{\beta}_b$ denote the image patches after replacing the original intensities with the multi-scale feature representations.

The advantage of using multi-scale feature representation in patch-based label fusion is shown in Fig. 2. Specifically, we examine the discriminative power of two target image points, designated by red '+' and red 'Δ' in Fig. 2. For clarity, we only use one atlas image in this example (bottom left of Fig. 2). The corresponding locations of the two target image points in the atlas image are designated with blue '+' and blue 'Δ', respectively. For each candidate point in the search neighborhood (i.e., blue dash boxes in Fig. 2), we compare the patch-wise intensity similarity w.r.t the target image point by using small-scale image patches ($3 \times 3 \times 3$), large-scale image patches ($17 \times 17 \times 17$), and our proposed multi-scale image patches, respectively. Figs. 2(a)–(c) shows the similarity maps obtained by comparing the target image patch and each candidate atlas image patch in the search neighborhood, where bright colors indicate high similarity, and dark colors indicate low similarity.

The principle behind patch-based label fusion methods is that two image patches should bear the same label if they have similar appearances. Therefore, the benefit of our multi-scale feature representation lies in its ability to recognize more reliable correspondences than the

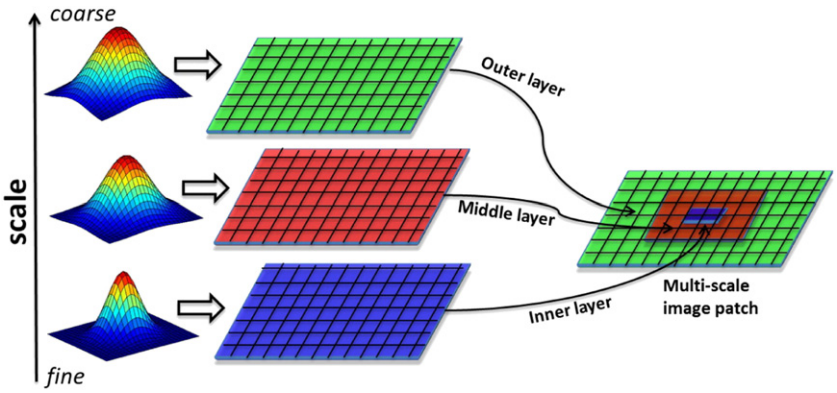

Fig. 1. Construction of the multi-scale image patch by adaptively replacing the intensity values with the convolved intensity values via multiple Gaussian filters.

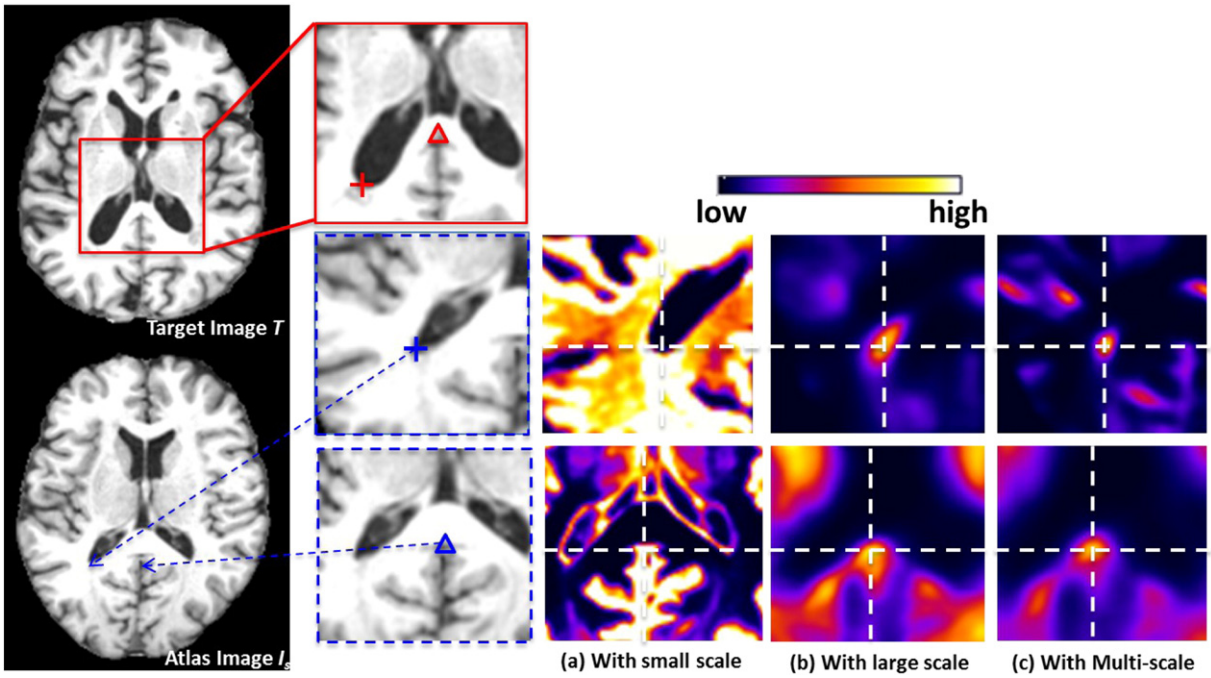

Fig. 2. The advantages of using the multi-scale patch representation in patch-based label fusion. Each marker ("+" or "Δ") denotes a pair of corresponding points in the atlas and target images, respectively. We show the advantages of using the multi-scale patch representation by examining the patch-wise similarity maps for a particular target image point w.r.t. each atlas image point. Compared with the similarity maps by using a small-scale image patch (a) and a large-scale image patch (b), our multi-scale image patch (c) can identify the corresponding locations in the atlas image domain more accurately. This can be noted by the fact that the similarity maps have high values (bright red) around the corresponding location and comparatively lower values (dark blue) elsewhere.

conventional image patches. As shown in Fig. 2(a), when using conventional small-scale image patches many image regions from obviously different anatomical structures present high similarities. This is because the appearance information in the small-scale image patch is too limited to characterize the complex anatomical structures. This has the undesirable effect of introducing misleading labels in label fusion. On the other hand, using conventional large-scale image patches can alleviate this issue by incorporating global information, but at the expense of losing discriminative power. This can be seen in Fig. 2(b), where a conventional large-scale image patch can approximately distinguish the atlas patches nearby the corresponding locations. However, a large number of atlas image patches belonging to different anatomical structures still present high similarities when using conventional large-scale image patches. Our multi-scale image patch combines both local and global information, which leads to a more reasonable similarity map as shown in Fig. 2(c). As we can see, our method can identify more accurate correspondences than using either small or large conventional patches. Thus, in the scenario of patch-based label fusion, the similarity map obtained by using our multi-scale image patch representation encourages assigning high weights to the true anatomical correspondences (with the correct labels) and also suppresses the atlas patches belonging to other structures (with incorrect labels).

Label-specific atlas patch partition

Since atlas image patches have label information, we can partition each atlas patch into a set of new label-specific atlas patches, thus separately encoding the image information for each individual label. Given the atlas image patch $\vec{\beta}_b$, we use $\vec{\gamma}_b$ to denote its associated label patch. Suppose there are M^b ($0 < M^b \leq M$) different labels in γ_b. Then, the proposed label-specific atlas patch set $\boldsymbol{P}_b$ consists of M^b label-specific atlas patches, i.e., $\boldsymbol{P}_b = \left\{\vec{p}_b^m | m = 1, \ldots, M^b\right\}$, where $\vec{p}_b^m$ is a column vector. Each element u in $\vec{p}_b^m$ preserve the intensity value $\vec{\beta}_b(u)$ if and only if $\vec{\gamma}_b(u)$ has the label l_m; otherwise, $\vec{p}_b^m(u) = 0$. Mathematically, we have $\vec{p}_b^m(u) = \vec{\beta}_b(u) \cdot \delta\left(\vec{\gamma}_b(u), l_m\right)$ and $\vec{\beta}_b = \cup_{m=1}^{M} \vec{p}_b^m$, where $\delta(.,.)$ is the same Dirac function as used in Eq. (2).

Fig. 3 demonstrates the construction of the label-specific atlas patch partition. For clarity, we only use the original 3×3 image patch in this example, instead of the above multi-scale image patch. Suppose that we have three atlas image patches and there are two labels (hippocampus and non-hippocampus) in each patch, i.e., $M^b = 2$ ($b = 1, 2, 3$). Next, for each atlas patch $\vec{\beta}_b$, we split it into two partial patches $\vec{p}_b^1$ and $\vec{p}_b^2$, which are denoted in blue (non-hippocampus) and red (hippocampus) in Fig. 3, respectively. Each label-specific atlas patch p_b^m preserves the intensity value only if the element bears the label l_m. Otherwise, use zero to represent the elements with a different label in the particular partial patch p_b^m.

Note that the number of image patches increases significantly after we partition each atlas patch into the label-specific atlas patch set. Thus, we propose to use the sparsity constraint in label fusion, in order to select only a small number of label-specific atlas patch $\vec{p}_b^m$ for representing the target image patch $\vec{\alpha}$. By replacing each conventional atlas patch with the label-specific atlas patches, the matrix of atlas patches $\boldsymbol{B}$ in Eq. (1) now expands to $\boldsymbol{P} = [\boldsymbol{P}_b]_{b=1,\ldots,Q}$. Then, the new energy function for label fusion can be reformulated as:

$$\hat{\vec{\xi}} = \arg\min_{\vec{\xi}} \frac{1}{2}\left\|\vec{\alpha} - \boldsymbol{P}\vec{\xi}\right\|^2 + \lambda\left\|\vec{\xi}\right\|_1, \text{s.t. } \vec{\xi} > 0, \tag{3}$$

where $\vec{\xi} = [\xi_b^m]$ is the weighting vector for each label-specific atlas patch $\vec{p}_b^m$. Since the goal of Eq. (3) is to minimize the difference between the target image patch and its sparse representation of label-specific atlas patches, the padded zero values in each label-specific atlas patch $\vec{p}_b^m$ have no influence when optimizing Eq. (3).

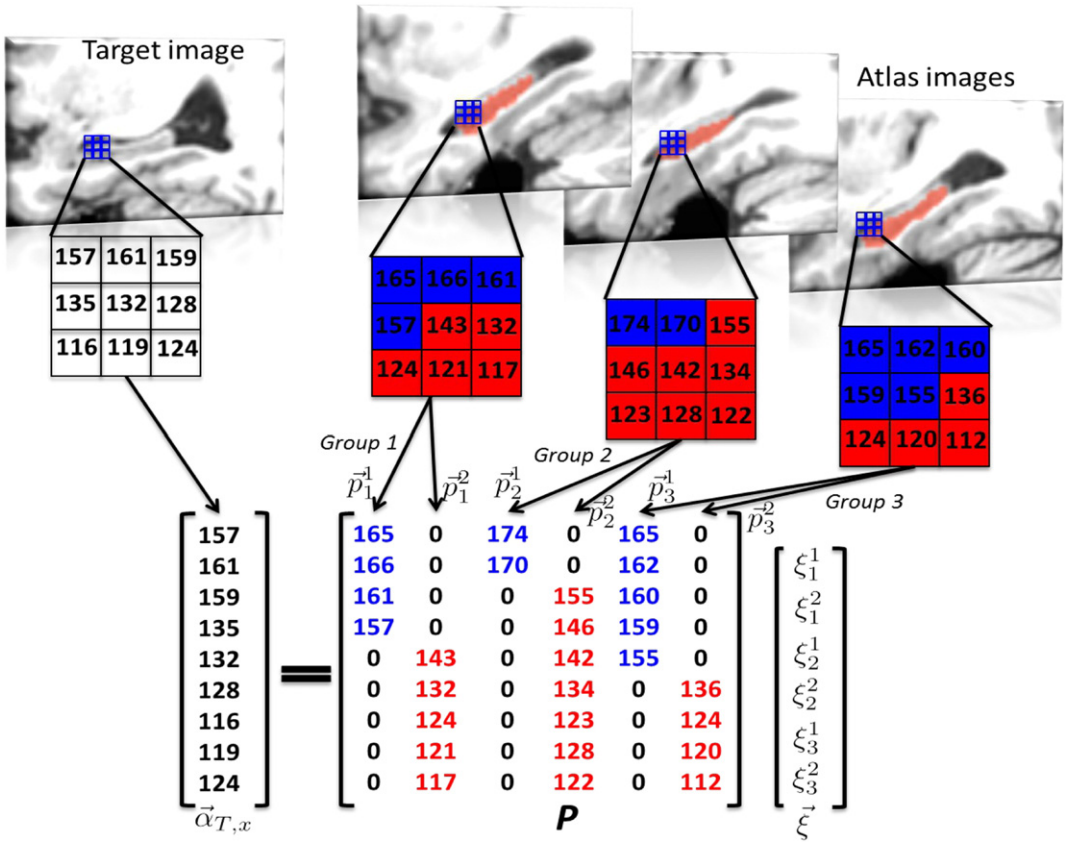

Fig. 3. The construction of a label-specific atlas patch partition.

Conventionally, each weight ξ_b^m in $\vec{\xi}$ is independently treated when optimizing Eq. (3). Here, we go one step further and enforce the group sparsity constraint on $\vec{\xi}$. Obviously, there are Q non-overlapping groups of label-specific atlas patches, where each group consists of a set of label-specific atlas patches split from β_b. Supposing that $\vec{\xi}_b = [\xi_b^m]_{m=1,\ldots,M^b}$ denotes the weights for all label-specific atlas patches within the original atlas image patch $\vec{\beta}_b$, the new energy function with group sparsity constraint is:

$$\hat{\vec{\xi}} = \arg\min_{\vec{\xi}} \frac{1}{2}\left\|\vec{\alpha} - \boldsymbol{P}\vec{\xi}\right\|^2 + \lambda_1 \sum_{b=1}^{Q}\left\|\vec{\xi}_b\right\|_2 + \lambda_2\left\|\vec{\xi}\right\|_1, \text{s.t. } \vec{\xi} > 0, \quad (4)$$

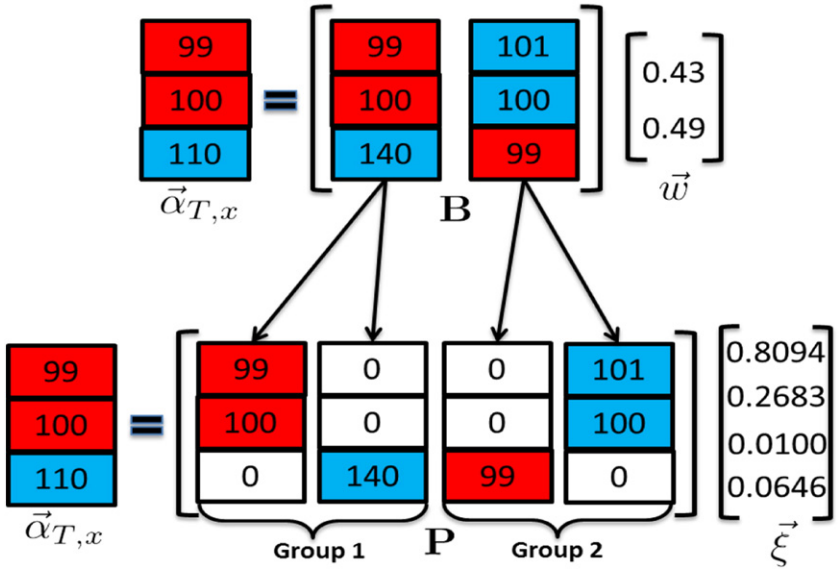

Fig. 4. The advantage of enforcing the group sparsity constraint in label fusion.

where λ_1 and λ_2 control the strength upon non-overlapping groups and the entire weighting vector $\vec{\xi}$, respectively. The new energy function falls into the scenario of sparse group LASSO (Friedman et al., 2010; Vincent and Hanse, 2014) which encourages sparsity *not only* for the entire weighting vector $\vec{\xi}$ (as reflected by the third term in Eq. (4)), *but also* for the number of selected groups (as reflected by the second term in Eq. (4)). The optimization of Eq. (4) can be efficiently solved by using the SLEP (Sparse Learning with Efficient Projection) software package (Liu et al., 2009b; Liu and Ye, 2010).

Since each $\vec{p}_b^{\,m}$ is only related with a particular label l_m, each element ξ_b^m in $\vec{\xi}$ represents the probability of labeling the center point x of the target image patch $\vec{\alpha}$ with the label l_m. Therefore, the labeling result on the target image point x can be obtained by:

$$\hat{L}_T(x) = \arg\max_{m=1,\ldots,M} \sum_{b=1}^{Q} \xi_b^m. \tag{5}$$

The advantage of using label-specific atlas patches is demonstrated by a toy example in Fig. 4, where we use red and blue to denote two different labels and we use numbers to represent the intensity values. For the sake of simplicity, we have only used two atlas patches in this example. Both the first atlas patch (i.e., the first column in $\boldsymbol{B}$) and the target patch $\vec{\alpha}$ belong to the same structure since their intensity values are in ascending order. If we estimate the weighting vector $\vec{w}$ based on the entire atlas patch by Eq. (1) ($\lambda = 0.01$), the weights for the first and second atlas patches are 0.43 and 0.49, respectively. According to Eq. (2), we have to assign the target point with the blue (incorrect) label. In our method, we first extend the matrix $\boldsymbol{B}$ to the label-specific atlas patch set $\boldsymbol{P}$, as shown in the bottom of Fig. 4, and then solve the new weighing vector $\vec{\xi}$ by Eq. (4) ($\lambda_1 = 0.1$ and $\lambda_2 = 0.1$). According to Eq. (5), the overall weights for red and blue labels are 0.8194 (0.8094 + 0.0100) and 0.3329 (0.2683 + 0.0646), respectively. Therefore, it is straightforward to correctly assign the target point with the red label. It is worth noting that, if only using the sparsity constraint on $\vec{\xi}$, the overall weights for the red and blue labels are 0.8850 (0.8000 + 0.0050) and 0.8004 (0.6901 + 0.1103), respectively. As we can see, the vote for the red label is only slightly better than the blue label. This example demonstrates the benefits of both the label-specific patch partitions and of enforcing group sparsity.

Hierarchical patch-based label fusion

In section Multi-scale feature representations, we have presented the use of multi-scale image patch to adaptively treat each element in the image patch. As observed in Fig. 2, combining global and local information can significantly increase the robustness and discriminative power of the image patch in label fusion. Along the same lines, we further propose to dynamically adjust the patch size from large to small during the label fusion procedure. The idea is to initially resort to global information (i.e., using a large patch size) to discard the misleading candidate atlas patches and then gradually use more local information (i.e., using smaller patch sizes) to refine the optimization of the energy function (in Eq. (4)) based on the remaining atlas patches.

In the beginning of patch-based label fusion, we propose to use a large patch size in order to capture the global image information. Since we use the sparsity constraint for solving the weighing vector $\vec{\xi}$, only a small number of image patches are selected to represent the target image patch $\vec{\alpha}$, since many elements in $\vec{\xi}$ are zero or almost zero. After discarding those unselected atlas patches, we can confidently reduce the patch size of those selected atlas patches and then repeat the whole label fusion procedure as described in sections Multi-scale feature representations and Label-specific atlas patch partition by using more detailed, local features. In this way, our label fusion method can iteratively improve the labeling results in a hierarchical way.

The advantage of our hierarchical patch-based label fusion is demonstrated in Fig. 5, where we aim to determine the label of the target image point x (red cross in Fig. 5) located near the boundary of the hippocampus (with ground-truth label corresponding to hippocampus). To determine the label for this target image point x, a set of candidate atlas image patches are examined in a $15 \times 15 \times 15$ search neighborhood (i.e., blue dash boxes). After patch pre-selection (Coupé et al., 2011), only around 2000 atlas patches are used in determining the label for target image point x. In order to explicitly show the advantage of our hierarchical patch-based label fusion scenario, we only use the original image patch, with neither multi-scale representation nor label-specific partition. Moreover, only the sparsity constraint is used to seek for the label fusion weights (i.e., Eq. (3)), instead of using the group sparsity constraint. In the first iteration, the patch size is set to $11 \times 11 \times 11$. In Fig. 5(a), we plot the sparse coefficients after solving Eq. (3). The red and blue plots correspond to the atlas patches with hippocampus label and non-hippocampus label, respectively. It is clear that a large number of atlas image patches with non-hippocampus labels are selected to represent the target image patch, which makes the selection of the underlying label somewhat arbitrary due to the fact that the overall weight for the non-hippocampus label is nearly the same as for the hippocampus. In the second iteration, we only focus on the remaining (selected) atlas patches by discarding the unselected atlas patches (with zero coefficients). At this point, we reduce the patch size from $11 \times 11 \times 11$ to $7 \times 7 \times 7$, in order to resort to the local image information to refine the sparse representation. Since a lot of misleading and noisy image patches have been removed, the task of sparse representation becomes relatively easier. As shown in Fig. 5(b), the overall weight voting for hippocampus dominates the weight for non-hippocampus. However, we still can see some large sparse coefficients for some non-hippocampus image patches. Thus, we finally repeat the same procedure with the patch size reduced to $3 \times 3 \times 3$. It can be observed in Fig. 5(c) that (1) only a few atlas patches are used to determine the label for the target image point x, and (2) all the selected atlas patches have the correct label w.r.t. the target image point x. It is worth noting that directly using the $3 \times 3 \times 3$ image patches from scratch does not result in good estimations, as indicated by the plot of sparse coefficients in Fig. 5(d). The main reason is that the appearance information from small image patches is too local to deal with the complex anatomical structures present in the brain images. On the contrary, our hierarchical label fusion framework uses the global image information to gradually remove the misleading candidate atlas patches, thus ensuring to obtain more accurate label fusion results when applying the small image patch size in the end.

Experiments

To evaluate label performance, the proposed label fusion method is compared to several existing state-of-the-art patch-based methods using publically available neuroimaging datasets. Specifically, the non-local weighting (Nonlocal-PBM) (Coupé et al., 2011; Rousseau et al., 2011), and the recently proposed sparse patch-based labeling method (Sparse-PBM) (Tong et al., 2012; Zhang et al., 2012) are tested. To assess label accuracy, the Dice ratio is used which measures the degree of overlap between two ROIs O_1 and O_2 as follows:

$$\mathrm{Dice}(O_1, O_2) = 2 \times \frac{|O_1 \cap O_2|}{|O_1| + |O_2|}, \tag{6}$$

where $|\cdot|$ means the volume of the particular ROI.

As shown in Table 1, an iterative process that uses varying configurations is implemented. In general, a configuration defines several partition layers defined within the patch. For instance, if the label fusion method initially starts with a $9 \times 9 \times 9$ patch it will be partition into

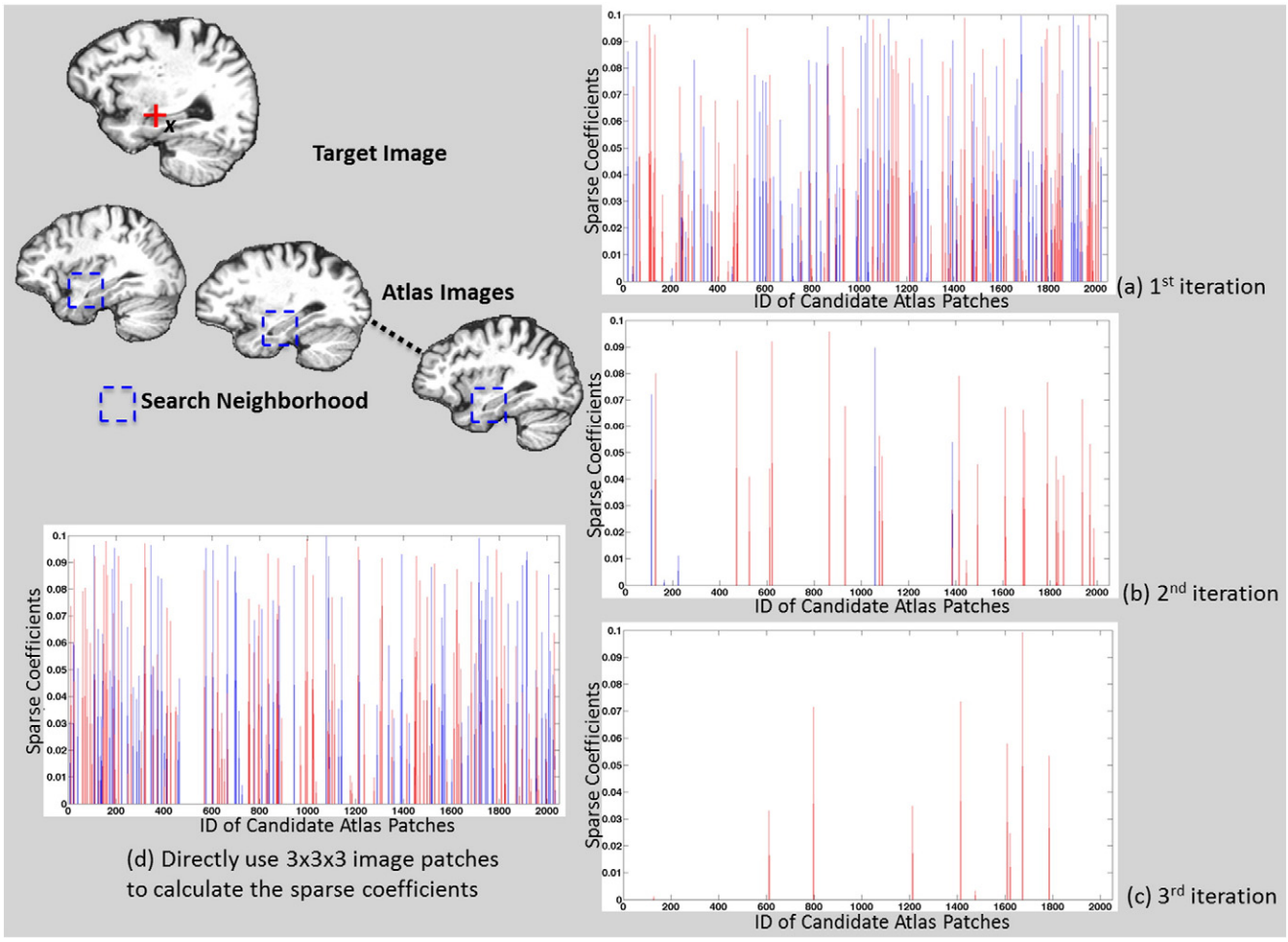

Fig. 5. The advantage of using hierarchical label fusion with dynamic patch sizes.

three layers. From the center point of the patch the [1,1,2] setting describes the width of each partition. For this particular setting, the width of the first two layers is 1 voxel and the width of the third layer is 2 voxels. Lastly, the [0.5,1,2] setting controls the kernel width that is used for smoothing. Likewise for this setting, 0.5 is the kernel width for the first layer, 1 is the kernel width for second layer, and 2 is the kernel width for the third layer. The above process is executed for two additional iterations that gradually reduce the patch size to $3 \times 3 \times 3$. These parameters are fixed throughout all the experiments. For the other counterpart methods, we report the results using the parameters that result in the best performance. Lastly, the values of λ_1 and λ_2 are both set to 0.1 for all experiments.

The remainder of this section is organized as follows: In Section 3.1 a comprehensive evaluation of the proposed label fusion method is performed using the ADNI dataset, then in Section 3.2 the parameters are fixed and the proposed method is used to segment the hippocampus in 7.0 T MR images. In section Experimental result of hippocampus labeling on the ADNI dataset, the proposed method is evaluated using the 14 mid-brain structures in the SATA MICCAI 2013 segmentation challenge dataset, and in section Experimental result on the LONI LPBA40 dataset the proposed method is evaluated using 54 manually labeled sub-cortical regions in the LONI LBPA 40 dataset (Shattuck et al., 2008). For a fair comparison, the Nonlocal-PBM and Sparse-PBM parameter settings reported in Coupé et al. (2011) and Zhang et al. (2012) respectively, are used.

Table 1
Example multiple layer configuration.

Patch size	Number of layers	Layer width	Gaussian Kernel size
$9 \times 9 \times 9$	3	[1,1,2]	[0.5,1.0,2.0]
$5 \times 5 \times 5$	2	[1,1]	[0.5,1.0]
$3 \times 3 \times 3$	1	[1]	[0.5]

Experimental result of hippocampus labeling on the ADNI dataset

In many neuroscience studies, accurate delineation of hippocampus is very important for quantifying the inter-subject anatomical difference and subtle intra-subject longitudinal changes, since the structural change of hippocampus is closely related with dementias, such as Alzheimer's disease (AD). In this experiment, we randomly select 23 normal control (NC) subjects, 22 MCI (Mild Cognitive Impairment) subjects, and 21 AD subjects from the ADNI dataset.[†] The following three pre-processing steps have been performed to all subject images: (1) Skull removal by a learning-based meta-algorithm (Shi et al., 2012); (2) N4-based bias field correction (Tustison et al., 2010); (3) intensity standardization to normalize the intensity range (Madabhushi and Udupa, 2006). Semi-automated hippocampal volumetry was carried out using a commercial high-dimensional brain mapping tools (Medtronic Surgical Navigation Technologies, Louisville, CO), which has been validated and compared to manual tracing of the hippocampus (Hsu et al., 2002). In this experiment, we regard the hippocampal segmentations from ADNI as the ground truth.

[†] http://adni.loni.ucla.edu/.

Table 2
Dice ratio mean, standard deviation, and mean computation time results for Nonlocal-PBM, Sparse-PBM, degraded versions of the proposed method, the proposed label fusion method and when used to label the hippocampus.

	Nonlocal-PBM	Sparse-PBM	Degraded_1	Degraded_2	Degraded_3	Proposed method
Dice ratio	86.6 ± 3.5	87.3 ± 3.4	87.9 ± 3.0	88.2 ± 2.5	87.6 ± 2.9	88.5 ± 2.2
Time (sec)	75	128	136	196	511	618

A leave-one-out strategy is used to compare the label performance of Nonlocal-PBM, Sparse-PBM, and proposed label fusion method. In each leave-one-out experiment, affine registration is first performed by FLIRT in the FSL toolbox (Smith et al., 2004) with 12 degrees of freedom and the default parameters (i.e., normalized mutual information similarity metric, and a search range of ± 20 mm in all directions). Then after the affine registration, a deformable registration is performed using the diffeomorphic demons (Vercauteren et al., 2009) method and the default registration parameters (i.e., smoothing sigma 1.8, and iterations in low, middle, and high resolutions as $20 \times 10 \times 5$).

To evaluate the contribution of each component in the proposed label fusion method, we compare our method with the three degraded versions of our method: *Degraded_1*: our method using only the multi-scale feature representation (with patch size of $9 \times 9 \times 9$), *Degraded_2*: our method using only the label-specific atlas patches (with patch size of $9 \times 9 \times 9$), and *Degraded_3*: our method using only the hierarchical labeling mechanism.

Using all 66 leave-one-out cases, the mean and standard deviation of the Dice ratios from the hippocampus label results are calculated and reported in Table 2. A few important observations can be made. Compared to the five other methods, the proposed label fusion method with no degradation achieves the highest Dice ratio results, obtaining approximately a 1.9% and 1.2% improvement over the Nonlocal-PBM and Sparse-PBM methods, respectively. Each component in the proposed label fusion method improves labeling accuracy as seen by the 0.6%, 0.9%, and 0.3% Dice ratio increases over Degraded_1, Degraded_2, and Degraded_3, tests respectively. The computation times by the 6 label fusion methods are also reported in the last row of Table 2. The computation environment of our experiments is 8 CPUs @ 3.0 GHz and 16 G RAM.

Since the improvement in label fusion is usually obtained around the boundary of hippocampus, it is interesting to examine the label results at the hippocampus surface. To perform this experiment we first construct ground-truth hippocampus surface mask and an estimated hippocampus surface mask. Then the distance at each vertex between two surfaces is computed. Table 3 shows the values of the averaged surface-to-surface distance and the maximum surface-to-surface distance by Nonlocal-PBM, Sparse-PBM, Degraded_1, Degraded_2, Degraded_3, and the proposed method with no degradation. We further perform the paired *t*-tests upon the surface distances. We observe that all degraded methods and the proposed method with no degradation have significant improvement ($p < 0.05$) over Nonlocal-PBM, the Degraded_1, Degraded_2, and the proposed method with no degradation have significant improvement ($p < 0.05$) over Sparse-PBM.

Experimental result on the 7.0 T MR images

With the advent of 7.0-Tesla MR imaging technology (Cho et al., 2010) the achievement of high signal-to-noise ratio (SNR), as well as a dramatic increase in tissue contrast compared to the 1.5- or 3.0-Tesla MR images, is possible. A visual comparison in provided in Fig. 6, which shows a typical brain image slice produced by a 7.0-Tesla scanner with resolution of $0.35 \times 0.35 \times 0.35$ mm^3 next to slice from a 1.5-Tesla scanner with a resolution of $1 \times 1 \times 1$ mm^3. These high-resolution images enable researchers to clearly observe fine brain structures with sub-milimetric precision. We believe that the 7.0-Tesla MR imaging technique has the potential to become the standard technique for discovering the morphological patterns in the human brain in the near future.

For the 7.0-Tesla scanner (Magnetom, Siemens), an optimized multichannel radiofrequency (RF) coil and a 3D fast low-angle shot (Spoiled FLASH) sequence were utilized, with TR = 50 ms, TE = 25 ms, flip angle 10°, pixel band width 30 Hz/pixel, field of view (FOV) 200 mm, matrix size $512 \times 576 \times 60$, 3/4 partial Fourier, and number of average (NEX) 1. The image resolution of the acquired images is isotropic, e.g., $0.35 \times 0.35 \times 0.35$ mm^3. The hippocampi were manually segmented by neurologists (Cho et al., 2010). All images were pre-processed by the following steps: 1) inhomogeneity correction using N4 bias correction (Tustison et al., 2010); 2) intensity normalization for making image contrast and luminance consistent across all subjects (Madabhushi and Udupa, 2006); 3) affine registration to the selected template by FSL.

Using 7.0-Tesla MR imaging technology, the proposed label fusion method is used to segment the hippocampus from twenty-one 7.0-Tesla MR brain images. Unfortunately, existing state-of-the-art deformable image registration methods that are developed for 1.5-Tesla or 3.0-Tesla MR images do not perform well when used on 7.0-Tesla MR images. In general, this is primarily due to the severe intensity inhomogeneity in 7.0-Tesla MR images, the richer texture information in 7.0-Tesla MR images (as seen in Fig. 6(b)), and that only a small segment of the brain covering the hippocampus is scanned, instead of the whole brain.

Since we have the manually labeled hippocampus for each 7.0-Tesla MR image, we can quantitatively measure label fusion accuracy using a leave-one-out cross validation strategy. The mean and standard deviation of the Dice ratios on hippocampus are (77.42 ± 3.44) % by Nonlocal-PBM, (79.29 ± 2.46) % by Sparse-PBM, and (82.65 ± 1.37) % by the proposed method. Furthermore, in Table 4 we list the average and maximum surface distances between the manually segmented and the automatically estimated hippocampus masks by three different label fusion methods. Fig. 7 shows the mappings of the surface distances on three typical 7.0-Tesla MR images.

Experimental result on the SATA MICCAI 2013 challenge dataset

Using the SATA dataset, provided by MICCAI 2013 segmentation challenge workshop (https://masi.vuse.vanderbilt.edu/workshop2013/index.php/Main_Page), 35 training samples (atlas images and labels) as well as a collection of 12 testing images are provided. There are 14 ROIs that cover accumbens area, amygdala, caudate, hippocampus, pallidum, putamen, and thalamus on both hemispheres. Since the organizers have provided all registered atlas images to each target

Table 3
Dice ratio mean, standard deviation, and maximum surface distance results when used to label the hippocampus (unit: mm). Symbols '$^+$' and '*' indicate significant improvement ($p < 0.05$) over the Nonlocal-PBM and sparse-PBM methods.

	Nonlocal-PBM	Sparse-PBM	Degraded_1^{+*}	Degraded_2^{+*}	Degraded_3$^+$	Proposed Method^{+*}
Mean	0.410 ± 0.15	0.380 ± 0.10	0.353 ± 0.10	0.342 ± 0.09	0.369 ± 0.12	0.334 ± 0.09
Max	4.359	3.742	3.317	3.000	3.464	2.450

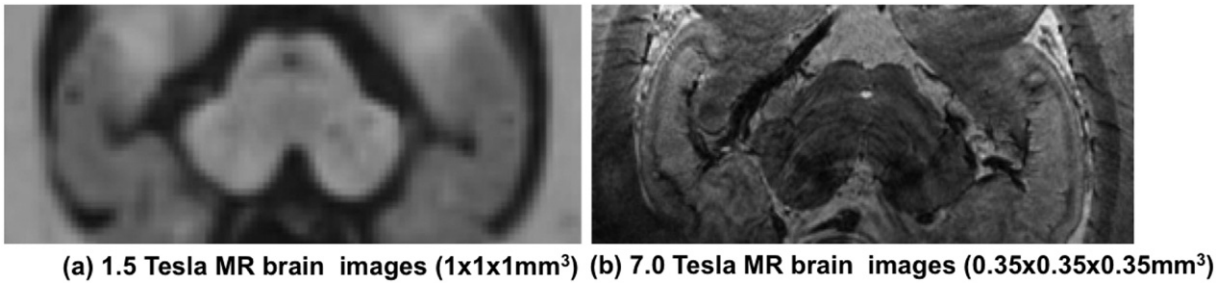

Fig. 6. The hippocampus shown by (a) 1.5-Tesla and (b) 7.0-Tesla MR scans. The 1.5-Tesla image has been enlarged to match the size of the 7.0-Tesla image for visual comparison purposes.

image to be labeled, no registration is needed for this experiment. After the proposed label fusion method generates the label results, they are submitted to the workshop organizer that returned the quantitative results shown in Table 5. The Dice ratios in all ROIs by the three label fusion methods are shown in Fig. 8. *It worth noting that the proposed method (named "UNC IDEA SuperMAS") is currently ranked the topmost label fusion method in this challenge* (http://masi.vuse.vanderbilt.edu/submission/leaderboard.html).

Experimental result on the LONI LPBA40 dataset

Here we evaluate the performance of label fusion using the LONI LBPA 40 dataset (Shattuck et al., 2008) that includes 40 brain images, and each brain image has 54 manually labeled ROIs. We randomly select 20 images as atlases and another 20 as target images. To label each target image, we first apply affine registration by FLIRT in the FSL toolbox (Smith et al., 2004) with 12 degrees of freedom and the default parameters (i.e., using the normalized mutual information similarity metric, and the search range $\pm$ 20 in all directions). Then after the affine registration, a deformable registration is performed using the diffeomorphic demons (Vercauteren et al., 2009) method and the default registration parameters (i.e., using the smoothing sigma 2.0, and iterations in low, middle, and high resolutions as $20 \times 10 \times 5$).

The Dice ratio mean and standard deviation measures for the 54 ROIs are provided in Table 6. The proposed method achieves a 3.15% and 1.51% improvement compared the Nonlocal-PBM and Sparse-PBM methods, respectively. Fig. 9 shows the Dice ratio in each ROI found by the Nonlocal-PBM (blue), Sparse-PBM (green), and the proposed method (red). The proposed label fusion method shows a significant improvement in 34 of 54 ROIs when compared to Nonlocal-PBM ('+' denoting significant improvement according to a paired t-test ($p < 0.05$)), and in 29 of 54 ROIs when compared to Sparse-PBM ('*' denoting significant improvement according to a paired t-test ($p < 0.05$)).

Comparison with other state-of-the-art methods on the IXI dataset

Recently, many multi-atlas based label fusion methods (Artaechevarria et al., 2009; Asman and Landman, 2012; Cardoso et al., 2013; Sabuncu et al., 2010a) have been developed to segment anatomical structures in medical images. STEPS (Similarity and Truth Estimation for Propagated Segmentations) (Cardoso et al., 2013) is one the most recent label fusion method that integrates image appearance information into the classic STAPLE algorithm (Warfield et al., 2004). Specifically, STEPS has achieved better segmentation results than other existing label fusion methods including those of Asman and Landman (2011), Asman and Landman (2012), Sabuncu et al. (2010b), and Yushkevich et al. (2010).

Here we compare segmentation performance of the proposed label fusion method with STEPS, Nonlocal-PBM, and Sparse-PBM using the IXI dataset (Hammers et al., 2003; Hammers et al., 2007) .[‡] The IXI dataset contains 30 subjects, each with 83 manually labeled ROIs. For the sake of comparison, we report the Dice ratios for the same 7 ROIs (Hippocampus, Amygdala, Caudate Nucleus, Nuc. Accumbens, Putamen,

Table 4
Dice ration mean, standard deviation, and maximum surface distance results found by Nonlocal-PBM, sparse-PBM, and the proposed label fusion method when used to label the hippocampus in 7.0-Tesla MR image (unit: mm).

	Nonlocal-PBM	Sparse-PBM	Our method
Mean	1.91 $\pm$ 0.41	1.43 $\pm$ 0.32	0.86 $\pm$ 0.16
Max	7.07	5.20	4.69

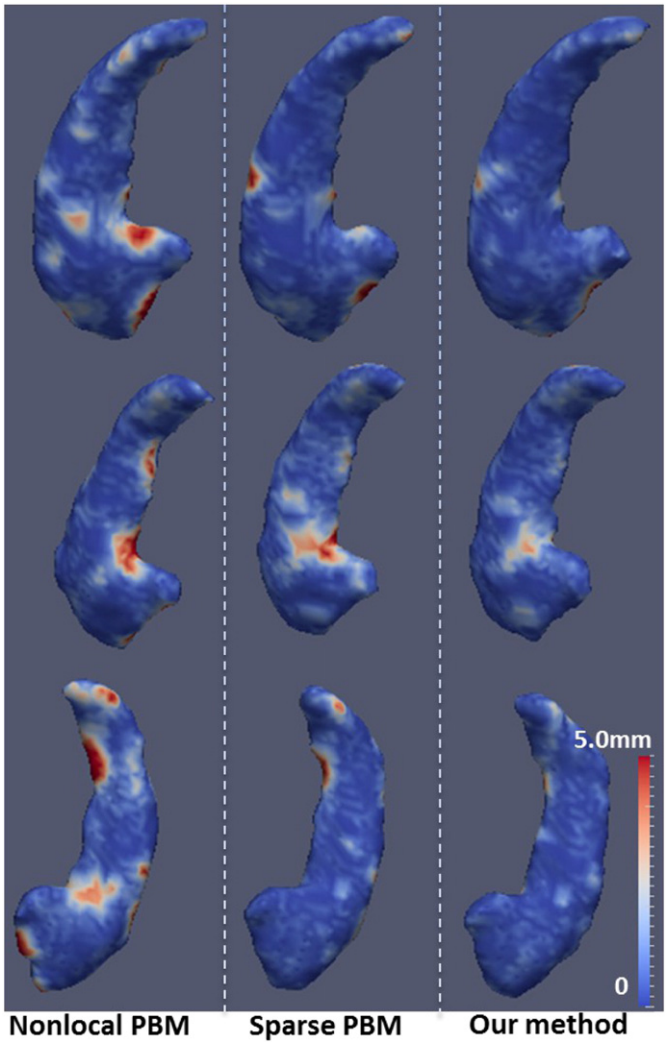

Fig. 7. Surface distance renderings obtained by Nonlocal-PBM, Sparse-PBM and our proposed label fusion method on 7.0-Tesla MR images.

[‡] The IXI dataset can be downloaded at http://biomedic.doc.ic.ac.uk/braindevelopment/index.php?n=Main.Datasets.

Table 5
Mean Dice ratio, standard deviation, median, maximum and minimum results found by Nonlocal-PBM, Sparse-PBM, and the proposed label fusion method using the SATA dataset.

	Mean	Standard deviation	Median	Max	Min
Nonlocal-PBM	85.81	2.80	86.95	89.04	80.61
Sparse-PBM	85.94	3.25	87.09	89.28	78.27
Proposed method	86.54	2.59	87.67	89.23	82.00

Table 6
Mean Dice ratio and standard deviation results found by Nonlocal-PBM, Sparse-PBM, and the proposed method using the LONI LPBA40 dataset.

	Nonlocal-PBM	Sparse-PBM	Proposed method
Mean and standard deviation	78.31 ± 3.52	79.95 ± 3.38	81.46 ± 2.25

Thalamus, Globus pallidus) originally reported in Cardoso et al. (2013). Similarly to Cardoso et al. (2013), we run Nonlocal-PBM, Sparse-PBM, and the proposed label fusion methods on all 30 subjects using a leave-one-out cross validation strategy. Table 7 shows the mean Dice ratio value for the 7 ROIs found by the different label fusion methods under test. As we can see, the proposed method achieves the best (i.e. greatest value) Dice ratio.

Discussion

Linear vs deformable image registration

In Rousseau et al. (2011), the authors propose the strategy that combines non-local label fusion with deformable image registration. According to their conclusions, accurate correspondences derived from deformable image registration could further improve non-local label fusion performance, especially when the intensity contrast is low. Since the overall goal of our paper is to improve labeling accuracy.

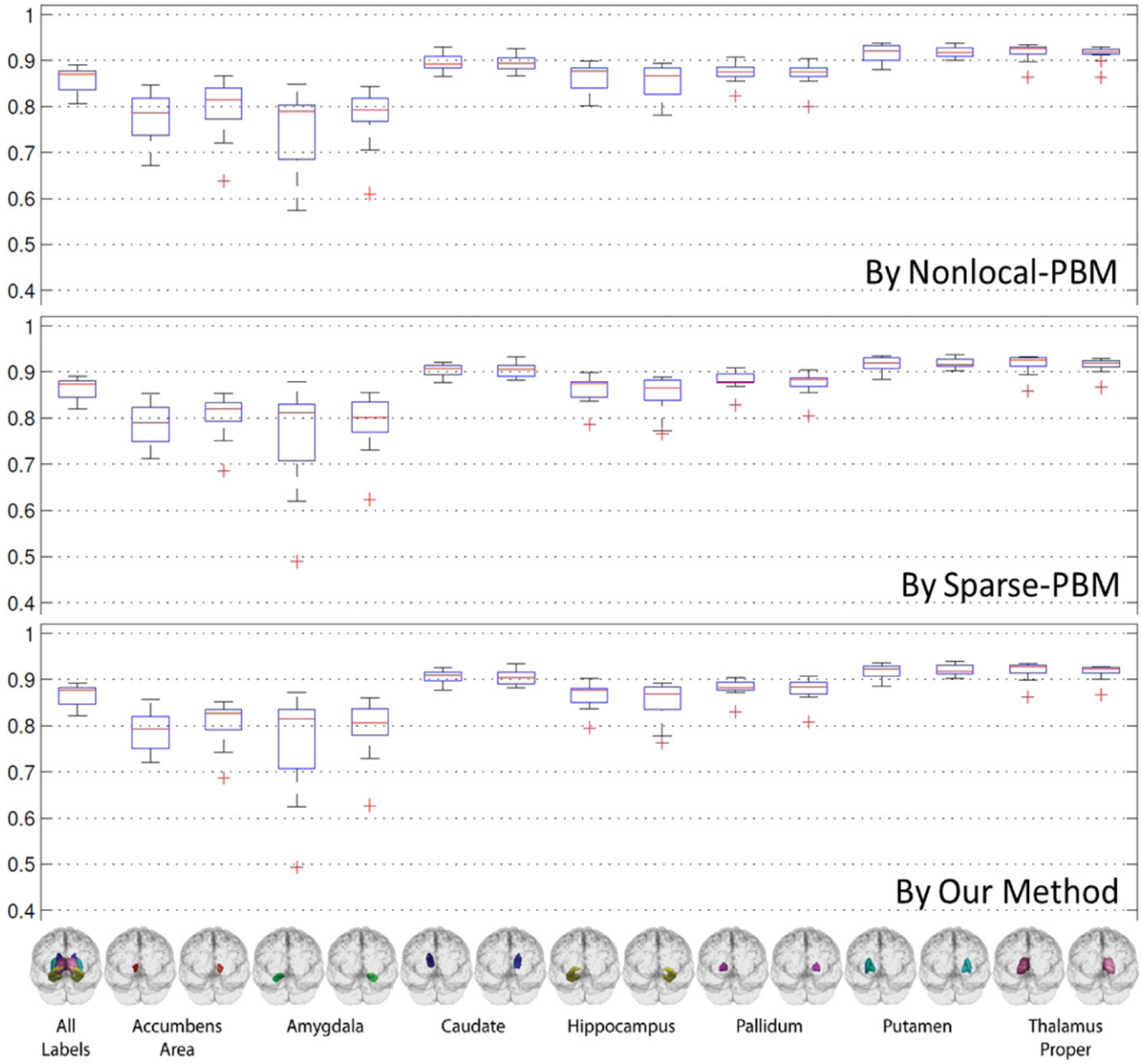

Fig. 8. Dice ratios for each ROI obtained by Nonlocal-PBM, Sparse-PBM, and the proposed label fusion method.

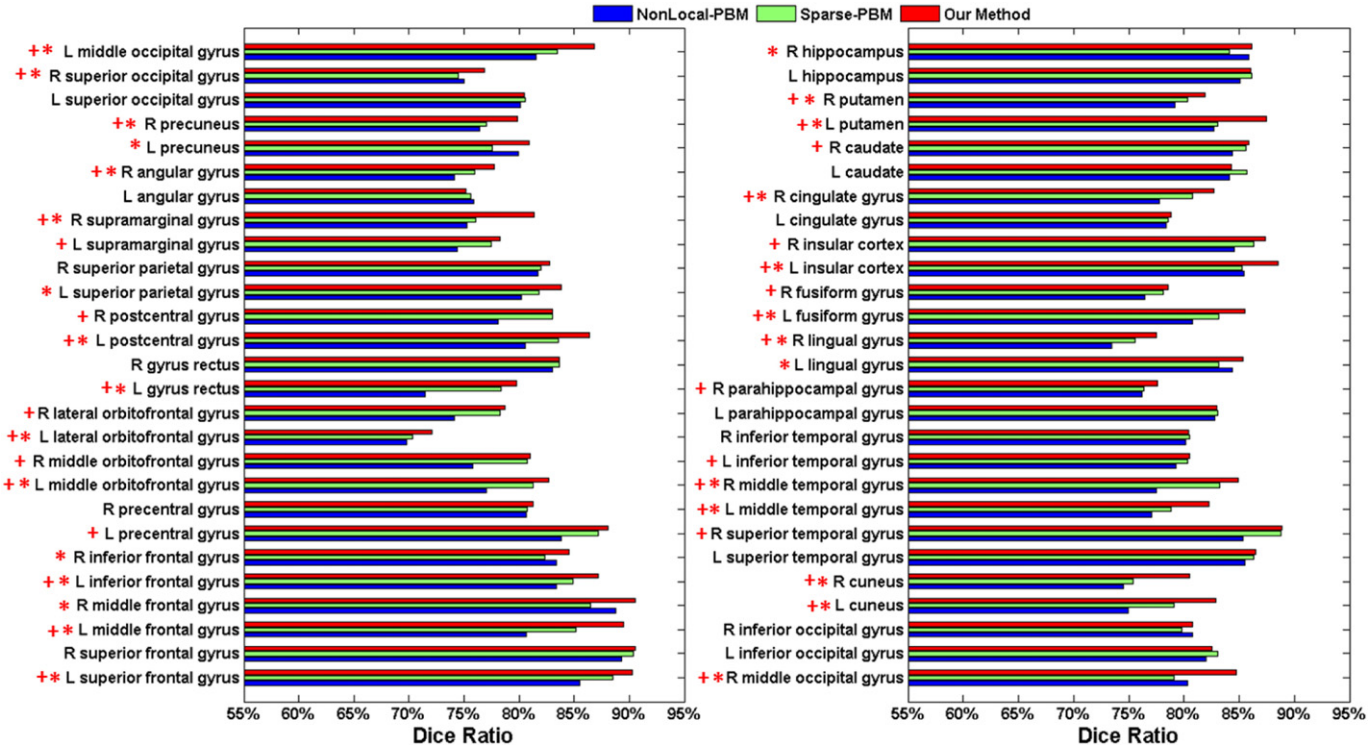

Fig. 9. Dice ratio for each ROI found by Nonlocal-PBM (blue), Sparse-PBM (green), and the proposed label fusion method (red). Symbols '+' and '*' indicate significant improvement ($p < 0.05$) with respect to Nonlocal-PBM and sparse-PBM, respectively.

In light of this, the proposed label fusion method was applied after deformable registration (using diffeomorphic demons) to map the labels from the atlas images to the target image. However, after several experiments we observed some interesting label fusion results that used linear registration instead of a non-linear one. In particular, Table 8 shows the mean and standard deviation of Dice ratios when segmenting the hippocampus using a 66 leave-one-out cross validation experiment. As shown in this table, we compared the Nonlocal-PBM, Sparse-PBM, Degraded_1, Degraded_2, Degraded_3, and our full label fusion method. Furthermore, the same dataset in section Experimental result of hippocampus labeling on the ADNI dataset was used with one exception: all the label fusion methods under test were executed after linear registration. Compared to the Dice ratios in Table 2, the segmentation of the proposed method is more accurate when linear registration is performed, and is less accurate when a deformable image registration is performed. Moreover, in hippocampus dataset, label fusion results after deformable image registration are more accurate than those after linear registration (87.9% by linear registration vs 88.5% by deformable registration), but at the expense of longer computational time (i.e., 20 min by linear registration vs 55 min by deformable registration).

Table 7
Mean Dice ratio results found by STEPS, Nonlocal-PBM, Sparse-PBM, and the proposed method using the IXI dataset.

	STEPS	Nonlocal-PBM	Sparse-PBM	Proposed method
Hippocampus	84.2	82.3	84.0	84.6
Amygdala	80.5	78.2	79.5	81.5
Caudate nucleus	89.2	88.5	88.9	89.5
Nuc. accumbens	69.5	68.9	69.1	70.6
Putamen	89.1	87.4	88.8	89.2
Thalamus	89.4	87.8	89.2	89.5
Globus pallidus	79.8	78.1	79.5	80.3

Overlapping vs non-overlapping layers in the multi-scale feature representation

In section Multi-scale feature representations, the image patch was partitioned into non-overlapping layers that may present blockness problems across different layers. In order to evaluate how this potential problem affects label fusion performance the two additional tests were evaluated: Include overlapping layers with a 1 image point overlap between two layers, and increase the number of layers in each image patch. For each additional test, the label fusion method was rerun with non-overlapping layers and with overlapping layers on the hippocampus dataset. After performing a 66 leave-one-out cross validation, the mean and standard deviation of the Dice ratios achieved by the proposed label fusion method, with non-overlapping layers, and with overlapping layers, were 87.91 ± 3.04 and 87.95 ± 2.96, respectively. Paired t-test indicates no significant statistical difference between when non-overlapping or overlapping layers are used. However, in our implementation there is a significant difference in computation time. Specifically, the time required when overlapping is used requires significantly more time when non-overlapping layers are used.

Limitations and future work

In order to efficiently obtain the multi-resolution feature representation at each point, we experimentally partition the image patch into several nested non-overlapping layers and assign each layer with a pre-determined Gaussian Kernel. However, as we demonstrated in our previous work (Wu et al., 2006b), each image point should have its own best scale to describe the local characteristics of the anatomical structure. Thus, one of our future works is to develop an adaptive method to use the best image patch size and the best set of smoothing kernels for each point. To further increase the computational efficiency of the proposed method, GPU processing using the CUDA programming technique can be used to exploit parallel patch operations. Lastly, the

Table 8
Dice ratio mean and standard deviation results when the hippocampus is labeled using only a linear registration.

	Nonlocal-PBM	Sparse-PBM	Degraded_1	Degraded_2	Degraded_3	Our method
Dice ratio	85.7 ± 4.0	86.2 ± 3.8	86.8 ± 3.0	87.2 ± 2.8	86.7 ± 3.3	87.9 ± 3.1

integration of the proposed label fusion method into an open-source stand-alone software package, like MARS (Multi-Atlas Robust Segmentation) that is hosted at NITRC (http://www.nitrc.org/projects/mars), would give other researchers direct access to the software developed in this manuscript.

Finally, although we address the limitation of existing label fusion methods that use fixed size image patches, many other works are aimed at improving label fusion performance from different perspectives. For example, Ta et al. (2014) introduced a new patch-based method using the 'PatchMatch' algorithm that provides competitive segmentation accuracy in near real-time. Results showed that their label fusion method can segment the hippocampus from MR images in less than 1 s. From the application point of view, the non-local based method has been adapted to multiple medical imaging studies, such as intracranial cavity extraction (Eskildsen et al., 2012; Manjón et al., 2014) and extraction of hippocampus structural features for early detection of AD (Coupé et al., 2012a; Coupé et al., 2012b). In our future work, we plan to evaluate our proposed label fusion method in other imaging-based studies (Chen et al., 2009; Liu et al., 2012; Verma et al., 2005).

Conclusion

In this paper, new techniques are used to improve multi-atlas patch-based label fusion performance. Specifically, each atlas patch is assigned a multi-scale feature representation; atlas image patches are partitioned into several label-specific patches based on existing label information; and a hierarchical label fusion mechanism that i\teratively improves the labeling result by gradually reducing patch size. Label fusion performance is evaluated using the ADNI dataset, 7.0-Tesla MR image dataset, SATA MICCAI 2013 segmentation challenge dataset, LPBA40, and IXI dataset. Compared to publicly available state-of-the-art label fusion methods, the proposed method has demonstrated the best label performance for each dataset. Lastly, it is worth noting that the proposed method has achieved the highest ranking in the SATA segmentation challenge.

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 4693) 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 Trans. Med. Imaging 28, 1266–1277.

Asman, A.J., Landman, B.A., 2011. Robust statistical label fusion through Consensus Level, Labeler Accuracy, and Truth Estimation (COLLATE). IEEE Trans. Med. Imaging 30, 1779–1794.

Asman, A.J., Landman, B.A., 2012. Formulating spatially varying performance in the statistical fusion framework. IEEE Trans. Med. Imaging 31, 1326–1336.

Cardoso, M.J., Leung, K., Modat, M., Keihaninejad, S., Cash, D., Barnes, J., Fox, N., Ourselin, S., 2013. STEPS: Similarity and Truth Estimation for Propagated Segmentations and its application to hippocampal segmentation and brain parcellation. Med. Image Anal. 17, 671–684.

Chen, Y., An, H., Zhu, H., Stone, T., Smith, J., Hall, C., Bullitt, E., Shen, D., Lin, W., 2009. White matter abnormalities revealed by diffusion tensor imaging in non-demented and demented HIV+ patients. NeuroImage 47, 1154–1162.

Cho, Z.-H., Han, J.-Y., Hwang, S.-I., Kim, D.-s, Kim, K.-N., Kim, N.-B., Kim, S.J., Chi, J.-G., Park, C.-W., Kim, Y.-B., 2010. Quantitative analysis of the hippocampus using images obtained from 7.0 T MRI. NeuroImage 49, 2134–2140.

Coupé, P., Eskildsen, S.F., Manjon, J.V., Fonov, V., Collins, D.L., 2012a. Simultaneous Segmentation and Grading of Anatomical Structures for Patient's Classification: Application to Alzheimer's Disease. NeuroImage 59, 3736–3747.

Coupé, P., Eskildsen, S.F., Manjon, J.V., Fonov, V., Pruessner, J.C., Allard, M., Collins, D.L., 2012b. Scoring by Nonlocal Image Patch Estimator for Early Detection of Alzheimer's Disease. NeuroImage: Clinical 1, 141–152.

Coupé, P., Manjon, 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, 940–954.

Eskildsen, S., Coupé, P., Fonov, V., Manjón, J., Leung, K., Guizard, N., Wassef, S., Ostergaard, L.R., Collins, D.L., 2012. BEaST: brain extraction based on nonlocal segmentation technique. NeuroImage 59, 2362–2373.

Friedman, J., Hastie, T., Tibshirani, R., 2010. A Note on the Group Lasso and a Sparse Group Lasso.

Hammers, A., Allom, R., Koepp, M., Free, S., Myers, R., Lemieux, L., Mitchell, T., Brooks, D., Duncan, J., 2003. Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. Hum. Brain Mapp. 19, 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. Hum. Brain Mapp. 28, 34–48.

Hsu, Y.-Y., Schuff, N., Du, A.-T., Mark, K., Zhu, X., Hardin, D., Weiner, M.W., 2002. Comparison of automated and manual MRI volumetry of hippocampus in normal aging and dementia. J. Magn. Reson. Imaging 16, 305–310.

Li, L.-J., Su, H., Xing, E.P., Li, F.-F., Hardin, D., Weiner, M.W., 2010. Object Bank: A High-Level Image Representation for Scene Classification and Semantic Feature Sparsification. Neural Information Processing Systems (NIPS).

Liu, J., Ye, J., 2010. Moreau–Yosida regularization for grouped tree structure learning. Advances in Neural Information Processing Systems.

Liu, J., Ji, S., Ye, J., 2009a. Multi-Task Feature Learning Via Efficient L2,1-Norm Minimization the Twenty-fifth Conference on Uncertainty in Artificial Intelligence Arlington. Virginia, USA.

Liu, J., Ji, S., Ye, J., 2009b. SLEP: Sparse Learning With Efficient Projections Software Manual. Arizona State University.

Liu, L., Zeng, L.-L., mail, Y.L., Ma, Q., Li, B., Shen, H., Hu, D., 2012. Altered Cerebellar Functional Connectivity with Intrinsic Connectivity Networks in Adults with Major Depressive Disorder PLoS ONE 7, e39516.

Liu, M., Zhang, D., Shen, D., 2012. Ensemble sparse classification of Alzheimer's disease. NeuroImage 60, 1106–1116.

Madabhushi, A., Udupa, J., 2006. New methods of MR image intensity standardization via generalized scale. Med. Phys. 33, 3426–3434.

Manjón, J.V., Eskildsen, S.F., Coupé, P., Romero, J.E., Collins, D.L., Robles, M., 2014. Nonlocal Intracranial Cavity Extraction. International Journal of Biomedical Imaging Article ID: 820205.

Rousseau, F., Habas, P.A., Studholme, C., 2011. A supervised patch-based approach for human brain labeling. IEEE Trans. Med. Imaging 30, 1852–1862.

Sabuncu, M.R., Yeo, B.T.T., Leemput, K.V., Fischl, B., Golland, P., 2010a. A generative model for image segmentation based on label fusion. IEEE Trans. Med. Imaging 29, 1714–1729.

Sabuncu, M.R., Yeo, B.T.T., Van Leemput, K., Fischl, B., Golland, P., 2010b. A generative model for image segmentation based on label fusion. IEEE Trans. Med. Imaging 29, 1714–1729.

Shattuck, D.W., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salamon, G., Narr, K.L., Poldrack, R.A., Bilder, R.M., Toga, A.W., 2008. Construction of a 3D probabilistic atlas of human cortical structures. NeuroImage 39, 1064–1080.

Shi, F., Wang, L., Dai, Y., Gilmore, J., Lin, W., Shen, D., 2012. LABEL: pediatric brain extraction using learning-based meta-algorithm. NeuroImage 62, 1975–1986.

Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E.J., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., Niazy, R.K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J.M., Matthews, P.M., 2004. Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage 23, S208–S219.

Ta, V.-T., Giraud, R., Collins, D.L., Coupé, P., 2014. Optimized PatchMatch for Near Real Time and Accurate Label Fusion. MICCAI 2014. Boston, USA.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B (Stat Methodol.) 58, 267–288.

Tong, T., Wolz, R., Hajnal, J.V., Rueckert, D., 2012. Segmentation of brain images via sparse patch representation. MICCAI Workshop on Sparsity Techniques in Medical Imaging, Nice, France.

Tong, T., Wolz, R., Coupé, P., Hajnal, J., Rueckert, D., 2013. Segmentation of MR images via discriminative dictionary learning and sparse coding: application to hippocampus labeling. NeuroImage 76, 11–23.

Tustison, N., Avants, B., Cook, P., Zheng, Y., Egan, A., Yushkevich, P., Gee, J., 2010. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging 29, 1310–1320.

Vercauteren, T., Pennec, X., Perchant, A., Ayache, N., 2009. Diffeomorphic demons: efficient non-parametric image registration. NeuroImage 45, S61–S72.

Verma, R., Mori, S., Shen, D., Yarowsky, P., Zhang, J., Davatzikos, C., 2005. Spatiotemporal maturation patterns of murine brain quantified by diffusion tensor MRI and deformation-based morphometry. Proc. Natl. Acad. Sci. U. S. A. 102, 6978–6983.

Vincent, M., Hanse, N.R., 2014. Sparse group lasso and high dimensional multinomial classification. Comput. Stat. Data Anal. 71, 771–786.

Wang, H., Suh, J.W., Pluta, J., Altinay, M., Yushkevich, P., 2011a. Optimal Weights for Multi-Atlas Label Fusion. Inf Process Med Imaging. 2011.

Wang, H., Suh, J.W., Pluta, J., Altinay, M., Yushkevich, P., 2011b. Regression-Based Label Fusion for Multi-Atlas Segmentation. CVPR 2011.
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, 903–921.
Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y., 2009. Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. 31, 210–227.
Wu, G., Shen, D., 2014. Hierarchical Label Fusion with Multiscale Feature Representation and Label-specific Patch Partition. MICCAI 2014. Boston, USA.
Wu, G., Qi, F., Shen, D., 2006. Learning-based deformable registration of MR brain images. IEEE Trans. Med. Imaging 25, 1145–1157.
Wu, G., Wang, Q., Zhang, D., Nie, F., Huang, H., Shen, D., 2014. A Generative Probability Model of Joint Label Fusion for Multi-Atlas Based Brain Segmentation. Medical Image Analysis 18, 881–890.
Yushkevich, P.A., Wang, H., Plutaa, John, 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, 1208–1224.
Zhang, D., Guo, Q., Wu, G., Shen, D., 2012. Sparse Patch-based Label Fusion for Multi-atlas Segmentation. MBIA, Nice, France.