Contents lists available at ScienceDirect

Medical Image Analysis

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

A transversal approach for patch-based label fusion via matrix completion

Gerard Sanroma[a], Guorong Wu[a], Yaozong Gao[a], Kim-Han Thung[a], Yanrong Guo[a], Dinggang Shen[a,b,*]

[a] *Department of Radiology and BRIC, University of North Carolina, Chapel Hill, USA*
[b] *Department of Brain and Cognitive Engineering, Korea University, Seoul, Republic of Korea*

ARTICLE INFO

Article history:
Received 22 October 2014
Revised 10 April 2015
Accepted 11 June 2015
Available online 20 June 2015

Keywords:
Label fusion
Matrix completion
Multiple-atlas segmentation

ABSTRACT

Recently, multi-atlas patch-based label fusion has received an increasing interest in the medical image segmentation field. After warping the anatomical labels from the atlas images to the target image by registration, label fusion is the key step to determine the latent label for each target image point. Two popular types of patch-based label fusion approaches are (1) *reconstruction-based approaches* that compute the target labels as a weighted average of atlas labels, where the weights are derived by reconstructing the target image patch using the atlas image patches; and (2) *classification-based approaches* that determine the target label as a mapping of the target image patch, where the mapping function is often learned using the atlas image patches and their corresponding labels. Both approaches have their advantages and limitations. In this paper, we propose a novel patch-based label fusion method to combine the above two types of approaches via matrix completion (and hence, we call it transversal). As we will show, our method overcomes the individual limitations of both reconstruction-based and classification-based approaches. Since the labeling confidences may vary across the target image points, we further propose a sequential labeling framework that first labels the highly confident points and then gradually labels more challenging points in an iterative manner, guided by the label information determined in the previous iterations. We demonstrate the performance of our novel label fusion method in segmenting the hippocampus in the ADNI dataset, subcortical and limbic structures in the LONI dataset, and mid-brain structures in the SATA dataset. We achieve more accurate segmentation results than both reconstruction-based and classification-based approaches. Our label fusion method is also ranked 1st in the online SATA Multi-Atlas Segmentation Challenge.

1. Introduction

Parcellation of the human brain structures is a key image processing step in many medical imaging studies related to computational anatomy and computer aided diagnosis (Li et al., 2014; Li et al., 2010; Nie et al., 2013; Nie et al., 2011). Manual annotation of anatomical structures is tedious and very time consuming, which makes it impractical in most of the current medical studies involving large amounts of imaging data. Therefore, high-throughput and accurate automated segmentation methods are highly desirable.

In the last two decades, multi-atlas segmentation (MAS) has emerged as a promising automated segmentation technique for segmenting a target image by propagating the labels from a set of annotated atlases. The use of multiple atlases makes MAS more capable of accommodating higher anatomical variability than using a single atlas. Moreover, as demonstrated in (Collins and Pruessner, 2009; Isgum et al., 2009; Rohlfing et al., 2004b), segmentation errors made by each individual atlas tend to be corrected when using multiple atlases. Generally, MAS consists of the following three steps: (1) the *atlas selection* step, where a subset of best atlases is first selected for a given target image based on a certain pre-defined measurement of anatomical similarity (Aljabar et al., 2009; Collins and Pruessner, 2009; Isgum et al., 2009; Rohlfing et al., 2004b; Sanroma et al., 2014a; Wu et al., 2007); (2) the *registration* step, where all selected atlases and their corresponding label maps are aligned to the target image (Klein et al., 2009; Shen and Davatzikos, 2002; Vercauteren et al., 2009; Wu et al., 2011); and finally (3) the *label fusion* step, where the registered label maps from the selected atlases are fused into a consensus label map for the target image (Artaechevarria et al., 2009; Cardoso et al., 2013; Coupe et al., 2011; Hao et al., 2013; Jia et al., 2012; Kim et al., 2013; Rousseau et al., 2011; Wang et al., 2011b; Warfield et al., 2004; Zikic et al., 2013). A great deal of attention has been put into the label fusion step, which is also the focus of the present paper, since it has a great influence on the final segmentation performance.

[*] Corresponding author.
E-mail address: dgshen@med.unc.edu (D. Shen).

http://dx.doi.org/10.1016/j.media.2015.06.002
1361-8415/© 2015 Elsevier B.V. All rights reserved.

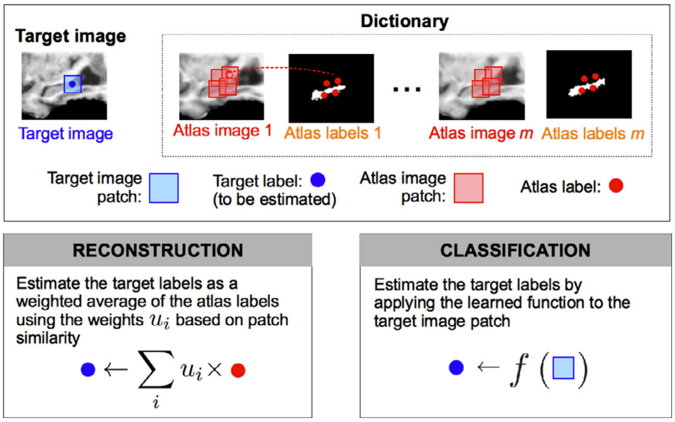
Dictionary
Target image
Target image
Atlas image 1
Atlas labels 1
Atlas image m
Atlas labels m
Target image patch:
Target label: (to be estimated)
Atlas image patch:
Atlas label:
RECONSTRUCTION
Estimate the target labels as a weighted average of the atlas labels using the weights u_i based on patch similarity
$\bullet \leftarrow \sum_i u_i \times \bullet$
CLASSIFICATION
Estimate the target labels by applying the learned function to the target image patch
$\bullet \leftarrow f(\square)$

Fig. 1. Illustration of reconstruction-based and classification-based label fusions. **Top**: a dictionary of atlas image patches (red squares) and their center labels (red circles) are used to estimate the target label (blue circle) in the center of the target image patch (blue square). **Bottom-left**: reconstruction-based approaches estimate the target label as a weighted average of the atlas labels, where atlas patches with higher similarity are assigned higher weights. **Bottom-right**: classification-based approaches estimate the target label by applying the relationships learned using the dictionary of atlas patches and labels. (See Sec. 3.1 for details about how the reconstruction and classification functions are computed.)

During the label fusion step, each target point is often independently labeled by using its own *dictionary* composed of the atlas patches and their labels selected from a neighborhood of the to-be-labeled target point (Coupe et al., 2011; Hao et al., 2013; Rousseau et al., 2011) (see the top panel in Fig. 1). Two recently popular label fusion approaches are the following: (1) reconstruction-based approaches, and (2) classification-based approaches. Reconstruction-based approaches are a particular type of weighted voting methods. As such, the target label is computed as a weighted average of the atlas labels (see the bottom-left panel in Fig. 1). Specifically, reconstruction-based approaches assign the weights based on the coefficients obtained by the linear reconstruction of the target patch using the dictionary of atlas patches (Tong et al., 2012; Zhang et al., 2012). This follows the idea of the image-similarity approaches, which assign higher weights to the atlas patches with more similarity to the target patch (Artaechevarria et al., 2009; Coupe et al., 2011; Rousseau et al., 2011). On the other hand, classification-based approaches use the dictionary of atlas image patches and their corresponding labels as the training set to learn the relationships between image appearance and anatomical labels (Hao et al., 2013) (Wang et al., 2011b). Then, in the labeling stage, the target label is estimated by directly applying the learned relationships to the target image patch (see the bottom-right panel in Fig. 1).

However, both reconstruction-based and classification-based approaches have their own limitations. Reconstruction-based approaches assume that the weights optimized based on patch-wise similarity are also optimal to fuse the labels. Unfortunately, as demonstrated in (Sanroma et al., 2014a), there is not always a clear relationship between appearance similarity and label consensus, and therefore similar atlas image patches could bear different labels. On the other hand, classification-based approaches overcome this limitation by specifically learning a mapping function from the image appearance domain to the label domain. However, all the atlas patches in the dictionary are given the same importance during the learning procedure, which may not be optimal since not all patches in the dictionary are equally representative for the target patch. Reconstruction-based approaches overcome this issue by adaptively weighting each atlas patch according to their estimated relevance in predicting the label of a particular target image point. In light of this, we present a novel label fusion method with the following contributions:

- We combine the advantages of both reconstruction-based and classification-based approaches by formulating label fusion as a matrix completion problem (but our method restricts to the *linear* sub-type of approaches). *First*, we build an *incomplete* matrix containing the target image patch as well as the atlas patches and their labels, where all the to-be-estimated target labels are missing. Based on the observation that there are high correlations among image patches and labels, we employ a low-rank constraint to estimate the missing elements in the above matrix. This entails taking full advantage of both row-wise and column-wise correlations (Candès and Recht, 2009), corresponding to the correlations in the vertical and horizontal directions of the matrix, respectively. As we will show, both reconstruction-based and classification-based approaches are particular cases where only row-wise (i.e., vertical) or column-wise (i.e., horizontal) correlations are exploited, respectively. By exploiting both types of correlations, our transversal method inherits the properties of both reconstruction-based and classification-based approaches, namely, (1) the property of the reconstruction-based approaches of representing the target patch as a weighted combination of the atlas patches, and (2) the discriminative ability of the classification-based methods in modeling the dependence of anatomical labels on the image appearance.
- We note that the labels at some parts of the image (e.g., deep inside the structures) can be determined more reliably than other parts (e.g., at boundaries of the structures), due to their anatomical characteristics and also due to their robustness to registration errors. However, most patch-based label fusion methods do not acknowledge this fact and label each target point independently. In this paper, we argue that it is more reasonable to let the high-confident points guide the labeling procedure of nearby less-confident points. Specifically, we embed our label fusion method into a sequential labeling framework that first labels the most confident target points and gradually labels those less-confident points iteratively. In this way, the anatomical labels estimated from the previous iterations can be used to help select more anatomically similar atlas patches to build the dictionary for improving the labeling.

We evaluate the label fusion performance of our proposed method on the ADNI, LONI, and SATA segmentation challenge datasets. We show that our proposed matrix completion based label fusion method outperforms both reconstruction-based and classification-based approaches. Moreover, we show that the sequential confidence-guided labeling scheme further improves our proposed method. Most importantly, our proposed method is ranked 1st in the online SATA Segmentation Challenge.

Note that a preliminary version of this work was presented in Sanroma et al., (2014b). The current paper (1) extends our previous work with the sequential confidence-guided labeling approach as described in Sec. 3.2, and (2) provides more exhaustive descriptions as well as illustrative examples of our extended method. We extensively (3) evaluate each component of our extended method by using additional datasets, and (4) compare it with the state-of-the-art methods.

2. Related work

With the advent of MAS, label fusion has become an increasingly active area of research. Label fusion is the key step that aims to segment the target image by finding a consensus among a set of registered atlas labels. The way in which the atlas information is used to derive the consensus segmentation has given rise to many different label fusion flavors. The simplest way, known as majority voting (MV), simply assigns each target point the label that appears most frequently among all corresponding atlas points (Heckemann et al., 2006; Rohlfing et al., 2005).

Another type of label fusion methods computes the target label as a weighted average of atlas labels, where weights are derived using local image similarity measurements. For example, local weighted voting (LWV) (Artaechevarria et al., 2009) is an example of this type of methods, which only uses the corresponding atlas points after registration to compute the label on each target point. Non-local weighted voting (Coupe et al., 2011; Rousseau et al., 2011) (NLWV) extends LWV by including all the atlas points within a small neighborhood, thus offering more flexibility to registration errors. Note that NLWV was originally inspired by image denoising ideas, where patches in the noisy image (i.e., target image) are *reconstructed* as a weighted average of patches in the database of images (i.e., atlas images). The only difference is that the label fusion methods reconstruct the target labels, instead of the target image. Motivated by the success of sparse representations in computer vision, sparse coding has also been studied in the context of label fusion (Tong et al., 2012; Zhang et al., 2012). The main idea is to reduce the number of contributing atlas labels to only a few relevant ones, thus offering better robustness to possible errors. The main idea behind all reconstruction-based methods is to first represent the target patch as a weighted combination of atlas patches, so that the target labels can be directly estimated using the ensemble of atlas labels according to the weights used in the representation.

On the other hand, Warfield et al. proposed a label fusion method, STAPLE (Rohlfing et al., 2004a; Warfield et al., 2004), that simultaneously estimates the target labels and the global performance of each atlas by means of the Expectation-Maximization algorithm (Dempster et al., 1977). Image appearance information has also been introduced into STAPLE to enhance the statistical modeling of the atlas performances. For example, non-local STAPLE (Asman and Landman, 2013) reformulates STAPLE to include priors based on the image similarity measurements. More recently, STEPS (Cardoso et al., 2013) introduces a local ranking strategy based on the image patch similarity into the STAPLE formulation.

Besides, there has been a wide interest in tackling the label fusion problem as a *classification* problem. In this case, the target label is computed as a function of the image features, where such a function models the dependency of atlas labels on the observed image patches. Different machine learning techniques have been used in this context of label fusion, such as support vector machines (Hao et al., 2013), polynomial regression (Wang et al., 2011b), random forests (Zhang et al., 2014; Zikic et al., 2013), and auto-context models (Kim et al., 2013). The main idea behind these methods is to learn a function that can discriminate among different possible labels based on the image appearance information.

Both reconstruction- and classification-based approaches follow a two-step approach, i.e., (1) the optimization step, where either the representation weights or the mapping functions are computed, and (2) the labeling step, where the target labels are estimated. Our method proposes a combination of reconstruction- and (linear) classification-based approaches by using matrix completion techniques (Candès and Recht, 2009), thus integrating the advantages of both approaches. Moreover, both optimization and labeling processes are carried out in a single step in our method.

However, in certain regions, the appearance information may be only *weakly* related to the underlying structural labels. In such case, it may be useful to rely on the putative anatomical information to reduce the ambiguity. For example, (Cardoso et al., 2013; Warfield et al., 2004) use Markov random fields (MRF) to enforce nearby target points to bear the same labels. Zhang et al. (Zhang et al., 2011) uses a similar assumption in a sequential labeling approach, where labels of more confident points are determined at earlier iterations and then the less confident points at later iterations are encouraged to bear the same labels as their neighboring more confident points.

Thus, we also embed our label fusion method into a sequential confidence-guided labeling framework by gradually labeling target points in decreasing order of confidence. However, instead of simply imposing neighboring points to bear the same labels, we use label information from previous iterations to select more meaningful atlas patches for labeling each target point.

3. Method

Our method is presented in two parts below. In Section 3.1, we present our label fusion method using matrix completion, and, in Section 3.2, we describe the sequential confidence-guided labeling framework. We denote images and label maps in bold capital letters, matrices in capital letters, vectors in lowercase letters with an overhead arrow, and scalars in lowercase letters.

3.1. Label fusion by matrix completion

3.1.1. Problem formulation

Suppose that we have a target image $\boldsymbol{T}$ and a set of m atlas images $\boldsymbol{A}_k$ along with their respective label maps $\boldsymbol{L}_k$ ($k = 1 \ldots m$), which have been already registered to the target image. The conventional label fusion approaches estimate the target label f at each voxel $x \in \Omega$ of target image $\boldsymbol{T}$ in a patch-wise manner. Let $\vec{t} \in \mathbb{R}^{p \times 1}$ denote a (column) vector containing the intensity values in the target image patch centered at voxel x, and matrix $A = [\vec{a}_1, \ldots, \vec{a}_i, \ldots, \vec{a}_n] \in \mathbb{R}^{p \times n}$ denote a dictionary of n candidate atlas image patches with the *highest similarity* to the target image patch in a search neighborhood of x (See Appendix A.3 for the details about building the dictionary). Following the same column order as the matrix A, $\vec{g} = [l_1, \ldots, l_i, \ldots, l_n]^\top \in \mathbb{R}^{n \times 1}$ is a (column) vector of ground-truth labels at the atlas patch centers, with each element $l_i \in \{-1, 1\}$ indicating either the absence (i.e., background) or the presence (i.e., foreground) of a given structure at the center of the respective atlas image patch $\vec{a}_i$.

As mentioned, label fusion can be regarded as a reconstruction or classification problem. As said, the reconstruction case is a particular type of weighted voting methods. As such, each target label f is computed as a linear combination of the atlas labels (Artaechevarria et al., 2009; Coupe et al., 2011; Rousseau et al., 2011; Zhang et al., 2012) as follows:

$$f = \vec{u}^\top \vec{g} \tag{1}$$

where $\vec{u} \in \mathbb{R}^{n \times 1}$ is a weighing vector to combine the atlas labels. (Note that, the resulting continuous label can be discretized to $\{-1,\ 1\}$ using the *sign* function). Weights in $\vec{u}$ encode the importance of each candidate atlas image patch in predicting the target label, and are computed so that the target patch $\vec{t}$ can be approximately reconstructed using the atlas patches in A (Tong et al., 2012; Zhang et al., 2012). This is,

$$\begin{aligned} &\min_{\vec{u}} \left\{ C_{\text{rec}}\left(\begin{bmatrix} \vec{t} \\ 1 \end{bmatrix}, \begin{bmatrix} A \\ \vec{1}^\top \end{bmatrix} \vec{u} \right) \right\} \Rightarrow \\ &\Rightarrow \begin{bmatrix} \vec{t} \\ 1 \end{bmatrix} \approx \begin{bmatrix} A \\ \vec{1}^\top \end{bmatrix} \vec{u} \end{aligned} \tag{2}$$

where $C_{\text{rec}}(\cdot)$ is the data fitting term penalizing reconstruction errors of the target patch. Note that the trailing 1's in the target and atlas patches encourage the weighting vector $\vec{u}$ to add up to one.

On the other hand, in the (linear) classification case, given a target image patch $\vec{t}$, its center label is determined based on the learned function, denoted as $\vec{v} \in \mathbb{R}^{p \times 1}$, aimed at mapping the appearance of the atlas image patch to its center label (Hao et al., 2013; Wang et al., 2011b). Assuming a linear function, the target label can be obtained by the following equation:

$$f = \begin{bmatrix} \vec{t} \\ 1 \end{bmatrix}^\top \vec{v} \tag{3}$$

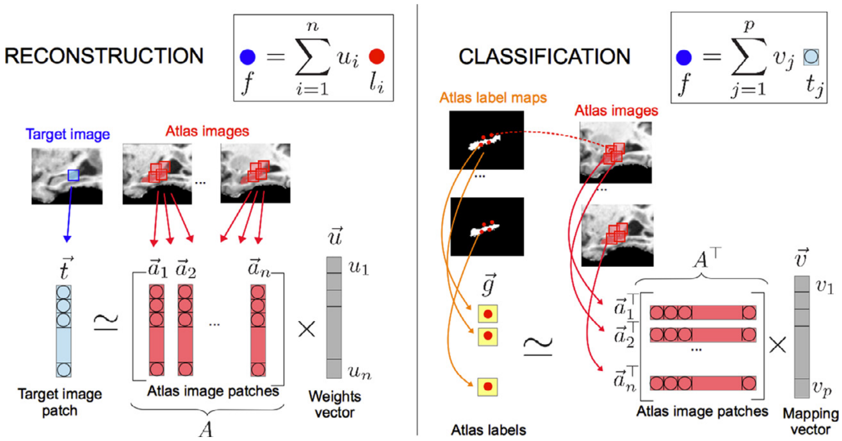

Fig. 2. Illustrative example of how the weights $\vec{u}$ and the mapping function $\vec{v}$ are computed in the reconstruction-based and classification-based approaches, respectively. *Note that the vectors of trailing ones have been omitted for simplicity.*

where the trailing 1 allows to include the bias term of the linear mapping in the last entry of $\vec{v}$ (as in the reconstruction case, the discrete label {−1,1} can be obtained using the *sign* function). The linear mapping function $\vec{v}$ encodes the relevance of each image feature in predicting the anatomical label and can be learned by minimizing the discrepancies between the predicted labels and ground-truth labels in the training set. This is,

$$\min_{\vec{v}} \left\{ C_{\text{cls}}\left(\vec{g}, \begin{bmatrix} A \\ \vec{1}^\top \end{bmatrix}^\top \vec{v}\right)\right\} \Rightarrow \\ \Rightarrow \vec{g} \approx \begin{bmatrix} A \\ \vec{1}^\top \end{bmatrix}^\top \vec{v} \qquad (4)$$

where $C_{\text{cls}}(\cdot)$ is a term penalizing the atlas label prediction errors (i.e., training errors). The procedures of reconstruction-based and classification-based methods are illustrated in Fig. 2.

3.1.2. Label fusion by matrix completion

We pose label fusion as a matrix completion problem, where the labels of to-be-estimated target patches are the missing entries in a specially constructed matrix. Furthermore, instead of predicting only the label at the center of each target patch, we also estimate all labels in the entire target image patch. Following the same order as in the atlas image matrix $A = [\vec{a}_1, \ldots, \vec{a}_i, \ldots, \vec{a}_n] \in \mathbb{R}^{p \times n}$, we arrange the label vector $\vec{l}_i$ of each atlas patch $\vec{a}_i$ into the atlas label matrix $L = [\vec{l}_1, \ldots, \vec{l}_i, \ldots, \vec{l}_n] \in \mathbb{R}^{p \times n}$.

Consider the four-quadrant matrix $Z = \begin{bmatrix} \begin{bmatrix} A \\ \vec{1}^\top \end{bmatrix} & \begin{bmatrix} \vec{t} \\ 1 \end{bmatrix} \\ L & \vec{f} \end{bmatrix}$, where each quadrant is a sub-matrix consisting of: (1) the atlas image matrix $A \in \mathbb{R}^{p \times n}$, (2) the atlas label matrix $L \in \mathbb{R}^{p \times n}$, (3) the target image patch $\vec{t} \in \mathbb{R}^{p \times 1}$, and (4) the to-be-estimated target label patch $\vec{f} \in \mathbb{R}^{p \times 1}$ (similarly defined as $\vec{l}_i$). The main idea of the reconstruction-based approaches implies that the target image patch can be expressed as a linear combination of the atlas image patches, whereas the main idea of the (linear) classification-based approaches implies that the label can be expressed as a linear combination of the image intensity values (with the vectors $\vec{u}$ and $\vec{v}$ in Fig. 2 containing the mixing coefficients in the reconstruction and classification cases, respectively). All these, in turn, imply that the four-quadrant matrix Z is highly correlated in both column-wise and row-wise fashions, and thus it is low-rank. We exploit this fact to recover the missing entries through rank minimization of the four-quadrant matrix (Candès and Recht, 2009). As we will see, this is equivalent to jointly using the properties of both reconstruction-based and classification-based approaches when estimating the target labels. In other words, we estimate the target labels by using *both* an ensemble of atlas labels *and* a learned discriminative function. Furthermore, by jointly estimating the labels for the whole target patch, we provide additional useful sources of correlation among the observed data to be leveraged by matrix completion. Fig. 3 illustrates the idea of our method.

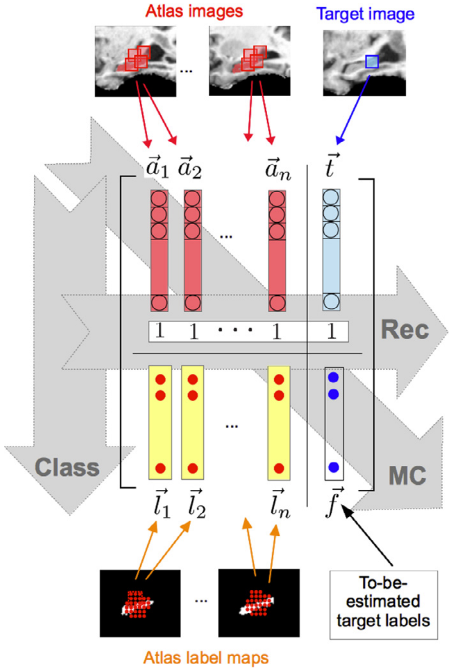
Atlas images
Target image
$\vec{a}_1$ $\vec{a}_2$ $\vec{a}_n$ $\vec{t}$
1 1 · · · 1 1
Rec
Class
MC
$\vec{l}_1$ $\vec{l}_2$ $\vec{l}_n$ $\vec{f}$
To-be-estimated target labels
Atlas label maps

Fig. 3. Each quadrant of the four-quadrant matrix is a sub-matrix, consisting of (1) stacked vectors of the atlas image patches (red), (2) stacked vectors of atlas label patches (yellow), (3) target image patch (light blue), and (4) to-be-estimated target labels (dark blue circles), respectively. Reconstruction-based methods utilize the correlations along the columns of the four-quadrant matrix, whereas classification-based methods utilize the correlations along the rows, as indicated by the horizontal and vertical shaded arrows, respectively. By imposing the low-rank constraint on this four-quadrant matrix, our method can simultaneously leverages the full row-wise and column-wise correlations for estimating the target labels, as indicated by the transversal shaded arrow.

3.1.3. Optimization

As denoted in Eq. (2), reconstruction-based methods assume that each target-patch column can be represented as a linear combination of all atlas-patch columns. On the other hand, as denoted in

Eq. (4), classification-based methods assume that each label-patch row can be represented as a linear combination of all image-patch rows. Such row- and column-wise dependences imply that the matrix Z is low-rank. This allows us to formulate the recovery of the missing target labels in $\vec{f}$ as a matrix rank minimization problem. By doing so, our method combines both reconstruction-based and classification-based methods, thus posing the recovery of target labels as a blend of row-wise and column-wise combinations. Since column-wise correlations describe the relationships between atlas image patches and target image patches, and row-wise correlations encode the dependence of the labels based upon the appearances of image patch, our MC-based label fusion method inherits the properties of both reconstruction-based and classification-based methods.

The key step in our approach is then finding the missing entries in $\vec{f}$ so that the rank of Z is minimized. Following Cabral et al. (2011), Goldberg et al. (2010), we seek a new matrix $\hat{Z}$ which satisfies the following conditions: (1) the rank of $\hat{Z}$ is low; and (2) the residue between the estimated and observed data in $\hat{Z}$ and Z is small. Due to the different natures of the two types of data in the matrix, we use two different cost functions to evaluate the residues: one for the image appearance, and another for the anatomical labels. Therefore, we define Θ_I and Θ_L as the sets of indices pointing to the entries in Z (i.e., pairs of row and column coordinates), corresponding to the *observed* image and label data, respectively (note that the indices of the to-be-estimated target labels in $\vec{f}$ are excluded from Θ_L). Accordingly, $z_{a,b}$, $(a, b) \in \Theta_I$, denotes the image-intensity value at position (a, b) in matrix Z (i.e., either red or blue quadrants of Fig. 3), and $z_{a,b}$, $(a, b) \in \Theta_L$, denotes the label value at position (a, b) in matrix Z (i.e., yellow quadrant of Fig. 3). The above objectives for finding the missing target labels can be formulated into the following convex optimization problem:

$$\min_{\hat{Z}} \left\{ \eta \|\hat{Z}\|_* + \frac{1}{|\Theta_1|} \sum_{(a,b) \in \Theta_I} c_I(\hat{z}_{a,b}, z_{a,b}) + \frac{\lambda}{|\Theta_2|} \sum_{(a,b) \in \Theta_L} c_L(\hat{z}_{a,b}, z_{a,b}) \right\} \tag{5}$$

where $\|\cdot\|_*$ denotes the nuclear norm (Candès and Recht, 2009) (i.e., the convex relaxation of the rank operator), $|\cdot|$ denotes the cardinality of the set, and $c_I(\cdot)$ and $c_L(\cdot)$ are the loss functions penalizing the estimation errors in the observed image and label entries, respectively. These two last terms follow the same idea as $C_{\text{rec}}(\cdot)$ and $C_{\text{cls}}(\cdot)$ of the reconstruction and classification approaches of Eqs. (2) and (4), respectively. We use the squared loss to penalize the error between the estimated image-intensity value $\hat{z}_{a,b}$ and the observed one $z_{a,b}$ ($(a, b) \in \Theta_I$), i.e., $c_I(\hat{z}_{a,b}, z_{a,b}) = (\hat{z}_{a,b} - z_{a,b})^2/2$, since it is suitable for the continuous values in the intensity images, and the logistic loss to penalize the label estimation errors, i.e., $c_L(\hat{z}_{a,b}, z_{a,b}) = \log(1 + \exp(-z_{a,b}\hat{z}_{a,b}))$, ($(a, b) \in \Theta_L$), since it is suitable for the binary values in the labels.

The first term in Eq. (5), which is controlled by the regularization parameter η, is responsible for decreasing the rank of the matrix $\hat{Z}$. Lower ranks tend to remove noisy variations in the matrix Z, thus improving the row-wise and column-wise correlations. This means that low rank minimization encourages each column to be represented as a linear combination of the other columns, and each row to be represented as a linear combination of the other rows, which correspond to the objectives of the reconstruction-based and classification-based approaches of Eqs. (2) and (4), respectively. Note that neither the weighting vector $\vec{u}$ nor the mapping function $\vec{v}$ are explicitly computed, as their computations are implicit in the minimization of the matrix rank. *The second term* in Eq. (5) is a feature error term which penalizes the discrepancy between the observed image data and the estimated image data in $\hat{Z}$. Having in mind that matrix $\hat{Z}$ is low-rank and thus contains significant column-wise correlations, this term encourages that the target patch is represented as a weighted average of the atlas patches, and then the atlas labels are transferred to the target by following the same representation. *The third term* in Eq. (5), which is controlled by the regularization parameter λ, is a label error term that penalizes the discrepancy between the ground-truth atlas labels and the estimated ones in the matrix $\hat{Z}$. Considering that matrix $\hat{Z}$ contains significant row-wise correlations, this term encourages that the dependencies between the atlas images and labels are effectively captured and, as consequence of the rank minimization, it also ensures that the missing target labels are filled-in following the same dependencies. We determine the values of η and λ empirically.

The optimization of Eq. (5) can be solved by an iterative algorithm that alternates between a gradient step and a shrinkage step (Goldberg et al., 2010). Specifically, in the gradient step, the matrix is updated so as to decrease the residual error, while, in the shrinkage step, the rank of the matrix is reduced. Since it is a convex optimization problem, the convergence to global optimum is guaranteed.

3.1.4. Summary

The matrix-completion based label fusion method can be represented as a function $\vec{f} = \text{MatComLF}(\vec{t}, A, L)$ that estimates the labels of a target patch in $\vec{f}$ using the target image patch in $\vec{t}$ and the dictionary of atlas image patches and labels in A and L, respectively. Since we estimate the label for the entire image patch and there are overlaps between image patches, we end up with multiple estimations for each target point. Accordingly, we first combine the multiple overlapping patch estimations into a continuous label map $\boldsymbol{F}$, as described in Appendix A.1, and then discretize the continuous label map to obtain the estimated target labels $\boldsymbol{D}$**,** as described in Appendix A.2. Table 1 shows a summary of our proposed algorithm for labeling an entire image.

3.2. Sequential confidence-based labeling

Selection of an appropriate dictionary is a key issue affecting the label fusion performance (Coupe et al., 2011; Hao et al., 2013). Recall that in $\vec{f} = \text{MatComLF}(\vec{t}, A, L)$, we obtain the dictionary (A, L) based on the image similarity between the target image patch and the neighboring atlas image patches (please refer to Appendix A.3). However, building the dictionary based solely on image similarity can undermine the label fusion performance, especially in challenging regions such as the boundaries of the structures, where similar atlas patches may bear different labels. To overcome this limitation we propose to use the prior knowledge about the labels on the target image to select the dictionary based on *both* image and label similarity with the target patch. Specifically, we adopt a sequential confidence-based labeling strategy where we first label the most confident target points (based on the magnitude of the continuous label values in $\boldsymbol{F}$) and then use this partial label information to refine the dictionaries used for labeling the less confident points at later iterations. As result, for each target patch $\vec{t}$, we obtain a refined dictionary $(\tilde{A}, \tilde{L}) \subset (A, L)$ containing a subset of atlas patches with both high image similarity and high label similarity. This process is summarized in Fig. 4.

3.2.1. Problem formulation

Assume that, at iteration s, we want to label a target image patch, denoted as $\vec{t}$, centered at x. We build the dictionary in a two-step approach: First, we build a dictionary of neighboring atlas image patches with high image similarity to the target image patch $\vec{t}$, denoted as $A = [\vec{a}_1, \ldots, \vec{a}_n]$ and $L = [\vec{l}_1, \ldots, \vec{l}_n]$. Next, we refine it based on the label similarity with the target label patch.

Let us denote the partial target label map from the previous iteration as $\boldsymbol{D}^{(s-1)}$. We extract the partial labels for the target patch at iteration $(s - 1)$, denoted as $\vec{d}$, consisting of a vector with entries in $\{-1, 1, \perp\}$, where -1, 1 and $\perp$ indicate background, foreground and unlabeled point, respectively. We then build the refined dictionary

Table 1
Algorithm for labeling one entire image using matrix-completion based label fusion.

```
Input: Target image $\boldsymbol{T}$, along with the atlas images and label maps $\boldsymbol{I}_k$ and $\boldsymbol{L}_k$, $k = 1 \ldots m$
Output: Estimated continuous and discrete target label maps $\boldsymbol{F}$ and $\boldsymbol{D}$, respectively
$F = \emptyset$ #set for aggregating the overlapping estimations
For Each voxel $x \in \Omega$ in the target image domain, do
        $\vec{t} = \mathrm{GetImgPatch}(\boldsymbol{T}, x)$
        $(A, L) = \mathrm{BuildDictionary}(\boldsymbol{I}_k, \boldsymbol{L}_k, \vec{t}, x),\ k = 1 \ldots m$ #see Appendix A.3
        $\vec{f} = \mathrm{MatComLF}(\vec{t}, A, L)$
        $F = F \cup \{\vec{f}\}$
End For
$\boldsymbol{F} = \mathrm{CombineOverlappingLabels}(F)$ #see Appendix A.1
$\boldsymbol{D} = \mathrm{Discretize}(\boldsymbol{F})$ #see Appendix A.2
```

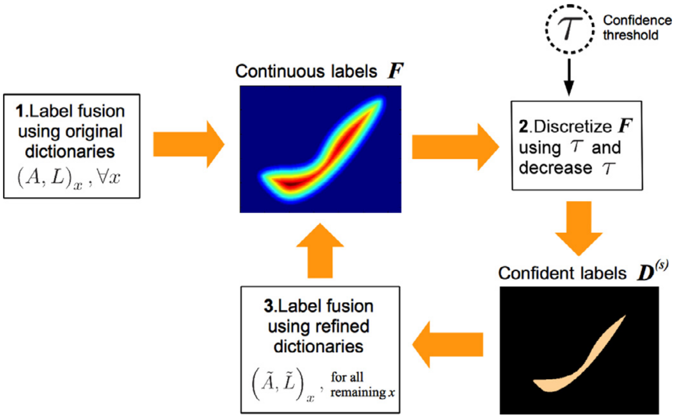
Confidence threshold
Continuous labels F
1.Label fusion using original dictionaries
(A, L)_x , ∀x
2.Discretize F using T and decrease T
Confident labels D^(s)
3.Label fusion using refined dictionaries
(Ã, L̃)_x , for all remaining x

Fig. 4. Overview of the sequential confidence-guided labeling framework: (1) We label each target point x using the original dictionaries, denoted as $(A, L)_x$. Note that, instead of obtaining a discrete label map, we obtain a continuous label map $\boldsymbol{F}$ indicating the label confidence values. (2) We obtain a partial segmentation, consisting of the most confident labels by discretizing the continuous labels using a pre-defined threshold τ, and then decrease the threshold. (3) We label the remaining unlabeled target points x by using the refined dictionaries $(\tilde{A}, \tilde{L})_x$ obtained with the help of the confident labels from previous iterations. *We repeat steps (2)–(3) until all the target points have been labeled.*

$(\tilde{A}, \tilde{L})$ using only the set of atlas patches with *high label similarity* to the partial target label patch, as defined in the following:

$$\left\{\vec{a}_i, \vec{l}_i \mid \mathrm{sim}\left(\vec{l}_i, \vec{d}\right) \geq \rho\right\} \tag{6}$$

where $0 \leq \rho \leq 1$ is a label similarity threshold and $\mathrm{sim}(\vec{l}_i, \vec{d})$ measures the similarity between the atlas label patch $\vec{l}_i$ and the partial target label patch $\vec{d}$. We define the label similarity measurement as the number of coincident labels in the atlas and target patches, normalized by the number labeled target points in the patch. More formally,

$$\mathrm{sim}(\vec{l}_i, \vec{d}) = |\mathrm{id}(\vec{l}_i) \cap \mathrm{id}(\vec{d})| / |\mathrm{id}(\vec{d})| \tag{7}$$

where $\mathrm{id}(\vec{d})$ is the indicator function denoting the set of indices in $\vec{d}$ containing foreground labels, and $|\cdot|$ denotes the cardinality of the set.

As result, the refined dictionary used to label the target image patch $\vec{t}$, denoted as $(\tilde{A}, \tilde{L})$, is composed by the atlas patches in the neighborhood of $\vec{t}$ satisfying both the image similarity criterion in Appendix A.3 (Eq. (A.2)) and the label similarity criterion of Eq. (6). Fig. 5 illustrates the dictionary refinement based on label similarity.

3.2.2. Summary

The whole iterative procedure is carried out as follows. At the first iteration, we compute the continuous label estimates $\boldsymbol{F}$ (which also represent the labeling confidence of the whole image) by using our proposed matrix-completion based label fusion method in Section 3.1. In the discretization step, we only assign labels to the most confident points according to a threshold τ, leaving the rest of points unlabeled. In the subsequent iterations, we re-compute the label confidences in the unlabeled parts by using the information of the labeled parts to refine the dictionary, as previously described. Note that we only need to re-compute the continuous labels in the unlabeled target points near to the labeled parts. In the end of each iteration, we discretize the new continuous estimates to obtain the partial label map $\boldsymbol{D}^{(s)}$, where we gradually decrease the confidence threshold τ across iterations. As result, we progressively estimate the labels for the less confident points with the guidance from labels of more confident points estimated in the previous iterations. This process has some similarity to simulated annealing (Sanroma et al., 2012a; 2012b), where a temperature parameter used to control the optimization process is gradually decreased and also the result from the previous iteration is used to initialize the next iteration. Fig. 6 shows an example of the evolution of the continuous label estimates across iterations, along with the resulting discrete confident labels. As we can see, the agreement of the continuous labels with the ground-truth labels improves across the iterations. In Table 2, we describe the algorithm of our full method.

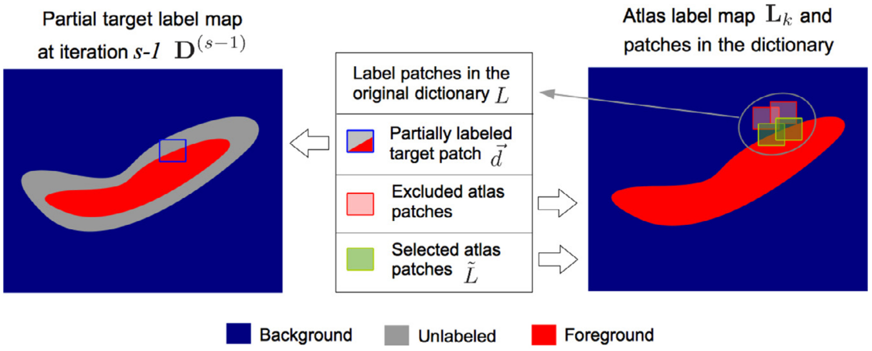
Partial target label map
at iteration s-1 D^(s-1)
Atlas label map L_k and
patches in the dictionary
Label patches in the original dictionary L
Partially labeled target patch d
Excluded atlas patches
Selected atlas patches L̃
Background
Unlabeled
Foreground

Fig. 5. Partially labeled target patch (i.e., blue square in the left-hand side). Atlas patches in the original dictionary (i.e., red and green squares in the right-hand side). We exclude the atlas patches in the dictionary with low label similarity to the partially labeled target patch (i.e., those red squares in the right-hand side).

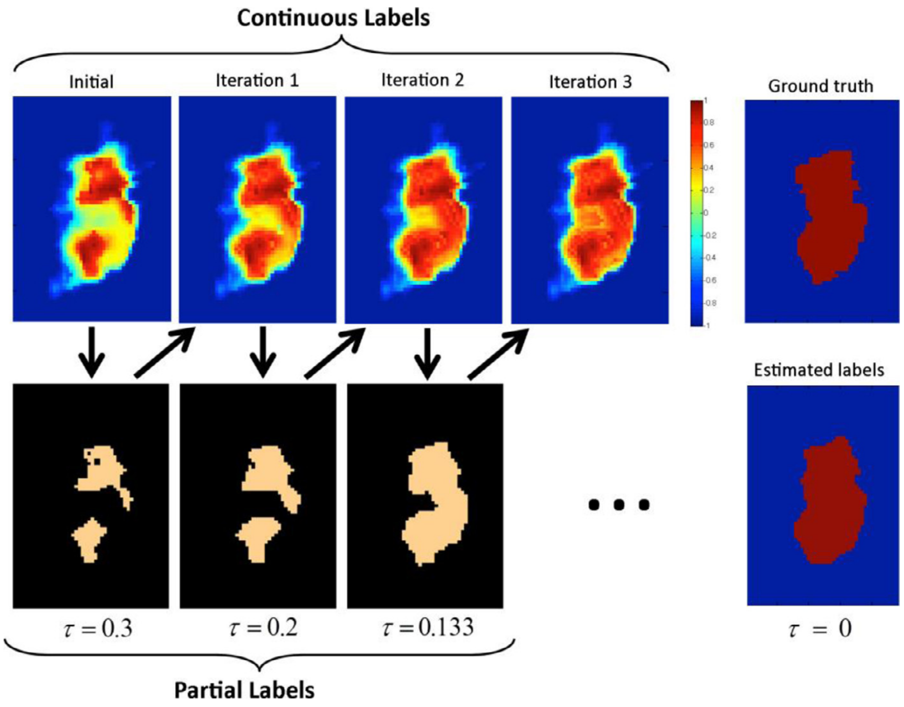

Fig. 6. Top-left: initial continuous label estimates of MC-based label fusion. Bottom-left: initial partial labels (with confidence threshold $\tau = 0.6$). Top-middle: evolution of the continuous label estimates across iterations using the information from confident labels. Bottom-middle: partial labels with the decreasing confidence threshold across iterations. Top-right: ground-truth target labels. Bottom-right: estimated target labels in the end of the sequential confidence-based labeling procedure.

Table 2
Algorithm of the sequential confidence-guided label fusion by matrix completion.

```
Input: Target image $\boldsymbol{T}$, atlas images and label maps $\boldsymbol{I}_k$ and $\boldsymbol{L}_k$, $k = 1 \ldots m$, initial discretization threshold $\tau_{\text{ini}}$, and patch selection threshold $\rho$
Output: Estimated target label map $\boldsymbol{D}$
$\boldsymbol{D}^{(0)}$ = Initializetounlabeled
$\tau = \tau_{\text{ini}}$
$s = 0$
While there are unlabeled points remaining in $\boldsymbol{D}^{(s)}$, do
        $s = s + 1$
        $F = \emptyset$ #set for aggregating the overlapping estimations
        For Each voxel $x \in \Omega$ in the target image $\boldsymbol{T}$, do
            $(\vec{t}, \vec{d})$ = GetImg&LabelPatch$(\boldsymbol{T}, \boldsymbol{D}^{(s-1)}, x)$
            $(A, L)$ = BuildDictionary$(\boldsymbol{I}_k, \boldsymbol{L}_k, \vec{t}, x)$, $k = 1 \ldots m$ #see Appendix A.3
            $(\tilde{A}, \tilde{L})$ = RefineDictionary$(A, L, \vec{d}, \rho)$ #Eq. (6)
            $\vec{f}$ = MatComLF$(\vec{t}, \tilde{A}, \tilde{L})$ #Section 3.1
            $F = F \cup \{\vec{f}\}$
        End For
        $\boldsymbol{F}$ = CombineOverlappingLabels$(F)$ #see Appendix A.1
        $\boldsymbol{D}^{(s)}$ = Discretize$(\boldsymbol{F}, \tau)$ #see Appendix A.2
        $\tau = \tau / \beta$ , $\beta \geq 1$
End While
$\boldsymbol{D} = \boldsymbol{D}^{(s)}$
```

4. Experiments

We evaluate the performance of the proposed method by conducting human brain anatomical segmentation experiments in a variety of datasets. In Section 4.2, we present hippocampus segmentation experiments in the ADNI[1] dataset. In Section 4.3, we segment all 16 subcortical and limbical structures in the LONI LPBA40[2] dataset (Shattuck et al., 2008). Finally, in Section 4.4, we provide segmentation results in the online SATA[3] Segmentation Challenge dataset, consisting of segmentations of 14 mid-brain structures.

In the ADNI and LONI-LPBA40 datasets, we conducted the following three pre-processing steps on all images before label fusion: (1) Skull stripping by a learning-based meta-algorithm (Shi et al., 2012); (2) N4-based bias field correction (Tustison et al., 2010); (3) ITK-based histogram matching for normalizing the intensity range. Prior to segmentation, we use FLIRT (Jenkinson et al., 2002) to perform linear (affine) alignment between each pair of images followed by non-rigid registration with diffeomorphic demons (Vercauteren et al., 2009). The images in the SATA dataset were already skull-stripped and the pairwise non-rigid registration was also performed.

4.1. Comparison methods

We compare our proposed label fusion method to a variety of related methods. As for the reconstruction-based methods, we compare

[1] http://www.adni-info.org/
[2] http://www.loni.usc.edu/atlases/Atlas_Detail.php?atlas_id=12
[3] https://masi.vuse.vanderbilt.edu/workshop2013/index.php/Main_Page

Table 3
Parameter values used in all the comparison methods.

Parameter	Details
Number of atlases m	In all the methods, we use the best $m = 15$ atlases selected according to mutual information, as this number of atlases has achieved an optimal performance in similar studies (Aljabar et al., 2009).
Patch size	We tried with isotropic patch sizes of 3, 5 and 7 voxels, and found that 5 yielded the best results.
Neighborhood radius ϵ	We tried with radius of $\epsilon = 1$, 2 and 3 and found that $\epsilon = 1$ performed the best in all the cases. By definition, we adopted the value of $\epsilon = 0$ for LWV.
LogReg and SPBL sparsity regularization α	We found that, the best amount of regularization for LogReg and SPBL was $\alpha_c = 0.5$ and $\alpha_r = 0.01$, respectively.
MCfull and MCdeg regularization parameters η, λ	We tried values in the range $\eta = 10^{-5} \ldots 1$ and $\lambda = 10^{-3} \ldots 10$, respectively, and we found that $\lambda = 0.05$ and $\eta = 10^{-4}$ yielded the best results in all datasets.
Label similarity threshold ρ	In the full version of our method (MCfull), we found $\rho = 0.9$ was the best value for the label similarity threshold, suggesting that enforcing high anatomical similarities in selection of atlas patches is beneficial for the segmentation performance.
Initial confidence threshold and decay parameter τ_{ini}, β	We set the value of the initial confidence threshold τ_{ini} according to the experiments in the beginning of Section 4.2. The decay parameter is fixed to $\beta = 1.5$.
STEPS (Cardoso et al., 2013)	There are three parameters to be tuned in STEPS, namely (1) the *kernel size* to measure image similarity in the local region (related to image patch size), (2) the *number of local labels*, and (3) the *amount of MRF regularization*. We tried with a range of values around the recommended values, and we kept the ones performing the best, which are the kernel size of 1.5, the number of local labels equal to 11, and the amount of MRF regularization equal to 4. Regarding MRF regularization, STEPS authors recommended a value in the range 0 . . . 5, which suggests that, in the present experiments, the MRF regularization has an important role for improving performance.

with Sparse Patch-based Labeling (**SPBL**) (Tong et al., 2012; Zhang et al., 2012) and some related image-similarity based methods such as Local Weighted Voting (**LWV**) (Artaechevarria et al., 2009) and Non-Local Weighted Voting (**NLWV**) (Coupe et al., 2011; Rousseau et al., 2011) label fusion. Note that the only difference between LWV and NLWV is the use of the neighborhood radius ϵ to build the dictionary, i.e., with $\epsilon > 0$ in NLWV while $\epsilon = 0$ in LWV. As for the classification-based methods, we have implemented a method that uses multi-task logistic regression (termed **LogReg**) for learning the mapping function between the appearance and the labels of Eq. (4). See Appendix B for more details.

In both reconstruction-based and classification-based approaches, we have tried either estimating only the center label for each target patch, or the whole patch. In order to keep the results as concise as possible, we only report the best estimation result (point-wise or patch-wise) for both reconstruction- and classification-based approaches. In most cases, we have found little difference between point-wise and patch-wise label estimation. Note that SPBL, NLWV and LogReg use exactly the same dictionary as our proposed method, thus providing fair comparison for these different label fusion methods.

We also compare with the state-of-the-art method **STEPS**[4] (Cardoso et al., 2013), which incorporates image similarity measurements into a statistical model of atlas performance to estimate the target labels. Moreover, it uses Markov Random field regularization to add spatial consistency by encouraging the neighboring target points to bear the same anatomical labels.

In order to assess each of our contributions, we further include two versions of our method in the comparison: (1) a degraded version (**MCdeg**) that uses only the matrix completion to label a target image in one-pass, as described in the algorithm of Table 1, and (2) the full version of our method (**MCfull**), as described in the algorithm of Table 2, which uses the sequential confidence-guided framework to refine the atlas dictionary.

Table 3 shows the values of the parameters used in all the comparison methods.

4.2. ADNI dataset

The ADNI dataset is provided by the Alzheimer's disease neuroimaging initiative and contains the segmentations of the left and right hippocampi which were obtained by a commercial brain mapping tool (Hsu et al., 2002). The size of each image is 256 × 256 × 256. We use 30 randomly selected subjects to test the performance of each of the segmentation methods. Due to the random selection, the prevalence of disease in our samples is similar to that in the original dataset, which is approximately 1/4 of Alzheimer's disease patients, 1/4 of healthy subjects, and 1/2 of subjects with mild cognitive impairment. In each segmentation experiment, one image is regarded as the target subject and the remaining 29 as the atlases. This process is repeated 30 times by regarding each image as target image once. Segmentation performance is assessed by the Dice ratio between manual annotations and automatic segmentations.

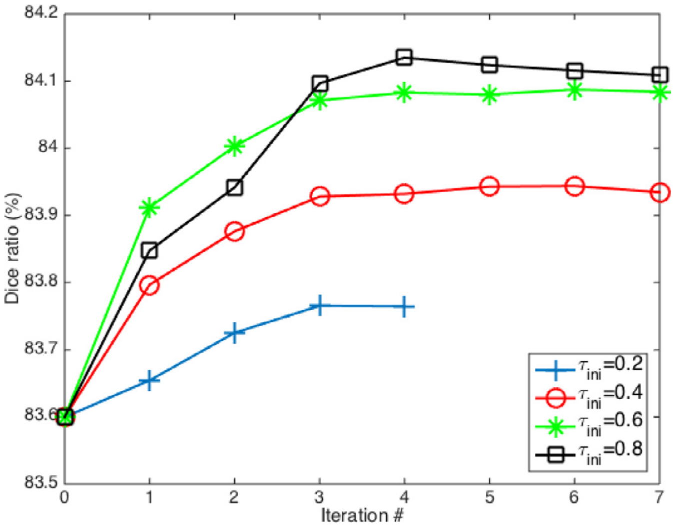

Fig. 7. Evolution of Dice ratio with iterations of MCfull for different values of τ_{ini}. Intermediate segmentation results at each iteration are obtained by thresholding the continuous label map at $\tau = 0$ in order to obtain a completely segmented image. Note that such completely labeled map is only used for obtaining the intermediate segmentation performance, while the partially labeled map according to the current value of the threshold is normally passed to the next iteration, as described in our method. Results at iteration 0 correspond to MCdeg, where the whole target image is labeled in one-pass without using any support from the high confident labels.

4.2.1. Sensitivity study to the initial confidence threshold

First of all, we evaluate the sensitivity of our method to the confidence threshold parameter τ_{ini}. In Fig. 7, we show the segmen-

[4] As part of the NiftySeg package downloadable from: `sourceforge.net/projects/niftyseg/`

Table 4
Dice ratio (%) in the ADNI dataset. We denote with markers * and + the statistically best and second best results among all the methods, respectively (according to a paired *t*-test at 5% significance level).

	STEPS	LWV	NLWV	SPBL	LogReg	MCdeg	MCfull
Left HC	81.46 (2.27)	81.26 (2.35)	82.42 (1.98)	82.61 (1.91)	82.83 (2.13)	83.56 (2.01) +	84.02 (2.15)*
Right HC	81.99 (2.72)	81.67 (2.93)	82.86 (2.35)	82.75 (2.21)	83.13 (2.42)	83.64 (2.36) +	84.15 (2.31)*
Overall	81.73 (2.50)	81.47 (2.64)	82.64 (2.17)	82.68 (2.05)	82.98 (2.26)	83.60 (2.17) +	84.08 (2.22)*

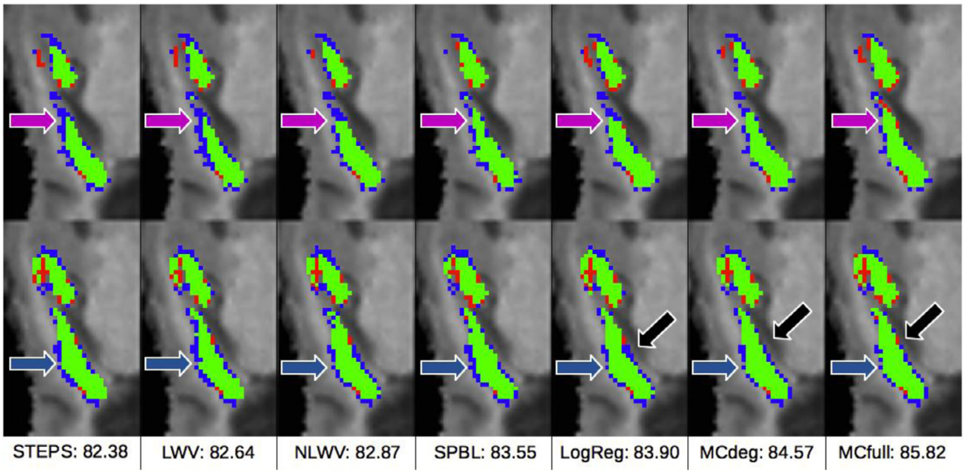

Fig. 8. Each column shows two consecutive slices of a typical example of hippocampus segmentation result by each method. Green labels denote coincidence between manual and automated segmentations (i.e., true positives), blue labels denote the parts of manually-segmented structures not detected by the automated method (i.e., false negatives), and red labels denote the parts of the automated segmentation that do not appear in the manual segmentation (i.e., false positives).

tation performance with iterations in the annealing procedure of MCfull, by using different initial values $\tau_{\text{ini}} = \{0.2, 0.4, 0.6, 0.8\}$ (See Appendix A.2 for details about the confidence-based discretization).

As we can see, segmentation performance increases w.r.t. the increase of the iteration number regardless the initial value of the confidence threshold, thus confirming the benefit of obtaining supports from the previously labeled points. Regarding the initial value of the threshold, results suggest to better start labeling only a few most confident points and then gradually labeling the rest of points in a supported way (i.e., $\tau_{\text{ini}} = 0.6$ and 0.8) rather than taking a higher risk at the beginning by labeling a large number of points in an unsupported way (i.e., $\tau_{\text{ini}} = 0.2$ and 0.4). On the other hand, using higher thresholds results in higher computational times because a larger number of points need to be considered at each iteration. The average computational times for completely labeling both left and right hippocampi in one subject for the confidence threshold values $\tau_{\text{ini}} = 0.2$, 0.4, 0.6 and 0.8 are 134, 214, 291 and 370 s, respectively[5]. In the case of $\tau_{\text{ini}} = 0.2$, the method usually completes the labeling of the subject before the 7th iteration. Taking into account both the performance and computational aspects, we choose the value $\tau_{\text{ini}} = 0.6$ in the rest of experiments.

4.2.2. Quantitative comparison

In Table 4, we show the segmentation performance by all the comparison methods for the left and right hippocampi (HC). Each value in the table shows the mean Dice ratio (and standard deviation) across 30 leave-one-out cross-validation experiments.

As we can see from these results, our proposed method (MCfull) achieves the best performance among all the methods, followed by the degraded version (MCdeg) which outperforms the rest of competing methods (according to a paired *t*-test at 5% significance level). Specifically, our proposed method (MCfull) outperforms both the reconstruction-based (LWV, NLWV and SPBL) and the classification-based (LogReg) approaches by ~1.5% and 1.1%, respectively. Regarding the degraded version of our method (MCdeg), we can see that it also outperforms both SPBL and LogReg by ~1% and ~0.6%, respectively, thus confirming the superiority of our combined, matrix-completion based approach, compared to the separate reconstruction-based and classification-based approaches. By comparing the results of the two versions of our method, we can see that the sequential confidence-guided framework provides a further improvement of ~0.5% with respect to MCdeg. Another interesting observation is that NLWV outperforms LWV by >1%, thus confirming the advantage of including neighboring atlas patches in label fusion as already noted by Rousseau et al. (2011). This has to be taken into account when interpreting the results of STEPS, which, like LWV, does not include the neighboring atlas patches in the dictionary. Thus, the ~0.3% performance improvement of STEPS over LWV is due to both the superior statistical estimation technique and the MRF-based regularization. Regarding the comparison of SPBL and LogReg, we observe that the classification-based approach outperforms the reconstruction-based approach by an average of ~0.3%. Each column in Fig. 8 shows two consecutive slices with the typical segmentation results by each comparison method.

The arrows point to the areas with the most significant differences among the methods. In general, the proposed methods, MCdeg and MCfull, show the highest true positives (green). Particularly, reconstruction-based methods tend to have more false negatives (blue). Comparing the results by STEPS and LWV, we can see that STEPS manages to reduce the false negatives in the area pointed by the purple arrow, probably due to the MRF regularization. LogReg obtains worse results than the proposed methods, MCdeg and MCfull, in the areas pointed by the black and purple arrows, respectively.

4.3. LONI dataset

The LONI LPBA40 dataset is provided by the Laboratory of NeuroImaging at UCLA and contains 40 brain images of size 220 × 220 × 184. Each image contains the annotations of 56 anatomical structures.

[5] Computational times of MATLAB/mex scripts on 4 Intel Core i7 CPUs at 2.5 GHz

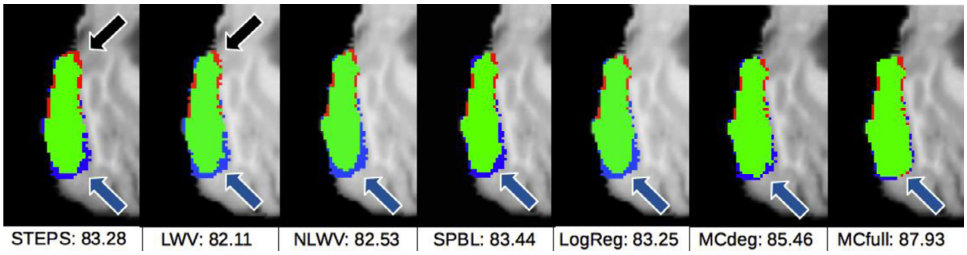

Fig. 9. Example segmentation results of the right gyrus rectus by all comparison methods. Green labels denote coincidence between manual and automated segmentations (i.e., true positives), blue labels denote the parts of the manually segmented structure not detected by the automated segmentation (i.e., false negatives), and red labels denote the parts of the automated segmentation that do not appear in manual segmentation (i.e., false positives).

Table 5
Dice ratio (%) in the LONI database. We denote with markers * and + the statistically best and second best results among all the methods, respectively (according to a paired *t*-test at 5% significance level). Note that we omit the marker* when no single method is statistically superior to the rest.

	STEPS	LWV	NLWV	SPBL	LogReg	MCdeg	MCfull
CN	82.10 (5.17)	82.24 (5.20)	82.99 (5.15)	83.64 (4.79)	83.29 (5.04)	83.77 (4.71) +	83.78 (4.67) +
GRe	78.63 (5.16)	78.14 (4.51)	78.81 (4.48)	79.13 (4.52)	79.19 (4.39)	79.67 (4.34) +	80.32 (4.74)*
HPC	82.73 (2.82)	82.60 (2.74)	83.33 2.69	83.63 (2.55)	83.51 (2.68)	83.69 (2.46) +	83.93 (2.57)*
PUT	83.22 (3.13)	82.44 (3.10)	82.92 3.10	84.83 (2.91)	83.73 (2.95)	84.42 (2.83) +	84.88 (2.96)*
LOG	71.10 (8.16) +	69.80 (8.05)	70.48 (8.23)	69.94 (8.23)	70.34 (8.40)	70.33 (8.43)	71.54 (7.80)*
PHG	79.76 (3.49)	79.23 (3.22)	80.11 (3.22)	80.33 (3.33)	80.26 (3.22)	80.64 (3.24) +	81.25 (3.53)*
IC	85.30 (2.34)	85.19 (2.15)	85.88 (2.12)	86.30 (2.07)	86.10 (2.10)	86.47 (2.07) +	86.55 (2.26) +
MOG	77.76 (6.42)*	76.94 (6.30)	77.46 (6.41) +	77.27 (6.37)	77.29 (6.44) +	77.29 (6.47) +	77.69 (6.50) +
Overall	80.07 (6.43)	79.57 (6.55)	80.25 (6.58)	80.51 (6.79)	80.46 (6.70)	80.79 (6.77) +	81.24 (6.53)*

We focus on the 16 subcortical and limbic structures, which consist of the left and right parts of the following structures: caudate nucleus (CN), gyrus rectus (GRe), hippocampus (HPC), putamen (PUT), lateral orbitofrontal gyrus (LOG), parahippocampal gyrus (PHG), insular cortex (IC), and middle orbitofrontal gyrus (MOG). As we did in the ADNI dataset, we compute the segmentation on each of the 40 images by using the remaining 39 as atlases, and this process is repeated for 40 times by leaving one different image out at each time. We assess the segmentation performance by using again the Dice ratio between manual annotations and automated segmentations by each method. In Table 5, we show the average Dice ratios (and standard deviations) across the 40 leave-one-out cross-validation experiments by each method in segmenting different structures.

Overall, our full method (MCfull) outperforms the rest of the methods, followed by our degraded method (MCdeg) according to a paired *t*-test at the 5% significance level. Specifically, it obtains average Dice ratio improvements of ~1% with respect to NLWV and 0.7% with respect to LogReg and SPBL. The degraded version of our method (MCdeg) obtains average improvements of >0.5% with respect to NLWV and >0.3% with respect to LogReg and SPBL, demonstrating the advantages of the combined approach over the reconstruction-based or classification-based approaches. Furthermore, MCfull achieves an improvement of >0.4% with respect to MCdeg due to the sequential confidence-guided framework. Results across different structures show that our full method achieves the best results in all the structures except for MOG, where STEPS obtains the best performance followed by our full method. The degraded version of our method (MCdeg) also outperforms the remaining methods in all the structures, except for the LOG and MOG, where MCdeg is outperformed by STEPS. Similarly, as we observed in the ADNI dataset, STEPS is superior to LWV, partly due to the benefits of using MRF regularization. Also, similarly as in the ADNI dataset, NLWV outperforms LWV by >1%, thus showing the advantages of including the neighboring atlas patches in the dictionary. Here, there are no significant performance differences between reconstruction- and classification-based approaches, as evidenced by the results of SPBL and LogReg, respectively. The benefit of the linear reconstruction strategy with sparsity constraint compared to the image similarity measurement is evidenced by differences in performance between SPBL and NLWV.

In Fig. 9, we further show one example of segmentation results of the right gyrus rectus by all the comparison methods.

Note the higher false negatives by all other methods in labeling the bottom part, except our proposed methods MCdeg and MCfull, as indicated by the blue regions pointed by the blue arrow. Both MCdeg and MCfull show improvement in this area with respect to other methods. Furthermore, MCfull shows the most accurate results, thus demonstrating the benefit of using the sequential confidence-guided framework. By comparing the segmentation results between STEPS and LWV as indicated by the black arrow, we can observe the increase in false positives perhaps due to the use of MRF regularization. We can also see that the MRF regularization is not able to correct the aforementioned false negatives as pointed by the blue arrow.

In order to give more insights on the performance of our full method, in Fig. 10, we further show the evolution of the segmentation performance with iterations. Similarly as in the ADNI dataset, we can see that the segmentation performance increases most significantly during the first 3 iterations, after which it stabilizes. The slight performance decrease at iterations $5-7$ (although not statistically significant) is possibly due to the fact that the newly labeled points at these iterations have lower confidence values and thus introduce some ambiguity. Recall that our 'annealing-like' approach uses the support of previously labeled points in a decreasing order of their confidence values. Therefore, points at the early iterations provide a more reliable support than points at the latest iterations. Note also that the minority of ambiguous points at the latest iterations cannot undermine the dramatic performance improvement achieved during the early iterations.

In order to give a visual insight of the proposed method, in Fig. 11 we also show the evolution of the continuous labels with iterations in labeling the right gyrus rectus. Our main purpose here is to show how the continuous label maps evolve with iterations after getting the supports from those confident labels of previous iterations.

The green circles denote the area where the initial estimate of the continuous label map highly disagrees with the manual segmentation. As we can see in the label maps of further iterations, our

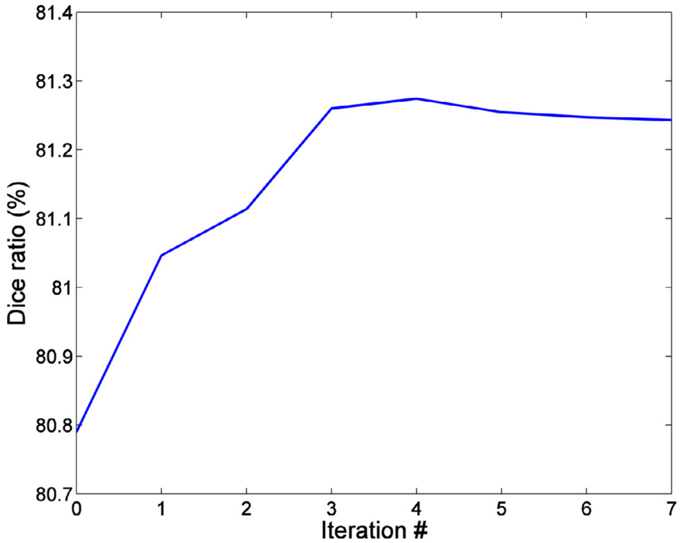

Fig. 10. Evolution of Dice ratio with iterations of our full method.

proposed method automatically corrects the disagreement in the mentioned area, while leaving the practically unchanged values for the rest of the (correct) areas.

4.4. SATA dataset

The SATA Segmentation Challenge Dataset is a publicly available dataset composed of 35 training and 12 testing brain MR images, respectively. Our main goal here is to evaluate the performance of our methods in an online challenge. Training images contain the manual annotations of 14 mid-brain structures, including the left and right parts of the accumbens area, amygdala, caudate, hippocampus, pallidum, putamen and thalamus proper. Testing images do not contain any label, so the estimated segmentations were submitted to the SATA Challenge website, where the performance statistics were computed and published in the leaderboard[6]. Pairwise non-rigid registrations between the images are also provided.

Note that one of the methods participating in the challenge, denoted as **PICSL**, is the ensemble method, composed of Joint Label Fusion method by (Wang et al., 2013) and the learning-based post-processing step for systematic error correction by (Wang et al., 2011a).

Table 6
Dice Ratio (%) and Hausdorff distance in the SATA challenge.

Method	Mean DR (std)	Mean HD (std)
MCfull	86.72 (2.83)	3.449 (0.650)
MCdeg	86.55 (2.88)	3.511 (0.718)
PICSL	86.43 (3.51)	3.458 (0.839)

In Table 6, we show the mean DR and the mean Hausdorff distance HD (in mm) obtained by the comparison methods.

As we can see, our proposed full method outperforms the rest of the methods in terms of both Dice ratio and Hausdorff distance. *It is worth noting that our proposed full method (MCfull) achieves the 1st position in the overall ranking*, whereas the degreaded version of our proposed full method (MCdeg) achieves the 3rd position (out of 14 methods). Specifically, our proposed full method obtains an improvement of ~0.3% with respect to the state-of-the-art ensemble method PICSL in both mean DR and HD, while having also lower standard deviations.

To give a further insight on the performance of our method, Fig. 12 shows the box plots across the different structures obtained by MCfull.

As we can see, the segmentation results of our method are quite accurate with the mean results on all the structures above 80%, and on some structures above 90%.

5. Conclusions

We have presented a novel label fusion method that combines the reconstruction-based and classification-based approaches by formulating label fusion as a matrix completion problem. Latent labels on the target image are regarded as the missing entries in a four-quadrant matrix, which are estimated by imposing the low-rank constraint. Furthermore, we have presented a sequential confidence-guided framework that gradually estimates labels at each iteration in decreasing order of confidence, while leveraging the support from the more confident labels of previous iterations. This reduces the ambiguity in the dictionary, thus leading to a significant performance improvement as confirmed by the experimental results. Our full method outperforms all other comparison methods in all the experiments presented. Also importantly, it outperforms all the methods listed in the website of the online SATA Segmentation Challenge (MICCAI 2013). The proposed matrix-completion based approach

[6] In the leaderboard, our methods are named “UNC MCseq” (MCfull) and “MCnoseq” (MCdeg), respectively. `masi.vuse.vanderbilt.edu/submission/leaderboard.html`

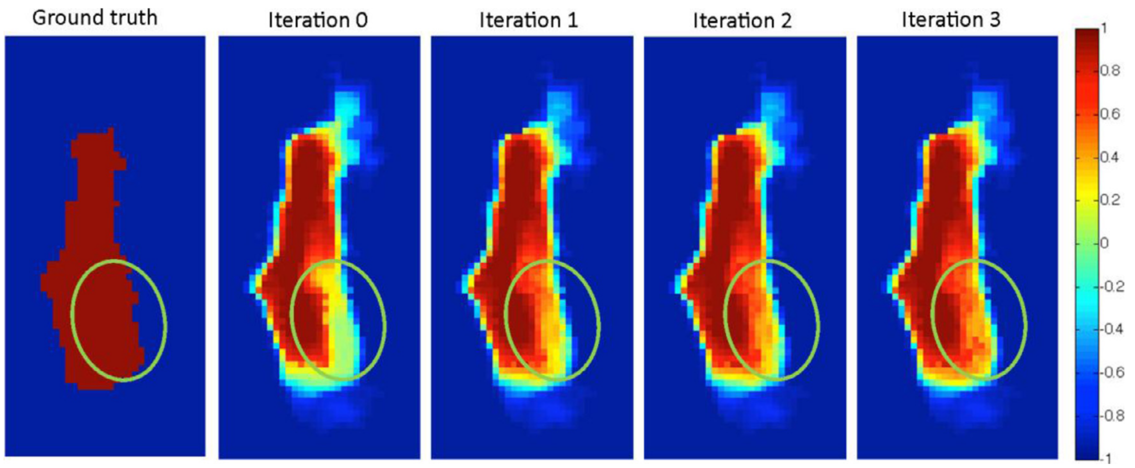

Fig. 11. From left to right: manual labels and the evolution of the continuous label maps obtained by the first 3 iterations of the sequential confidence-guided framework.

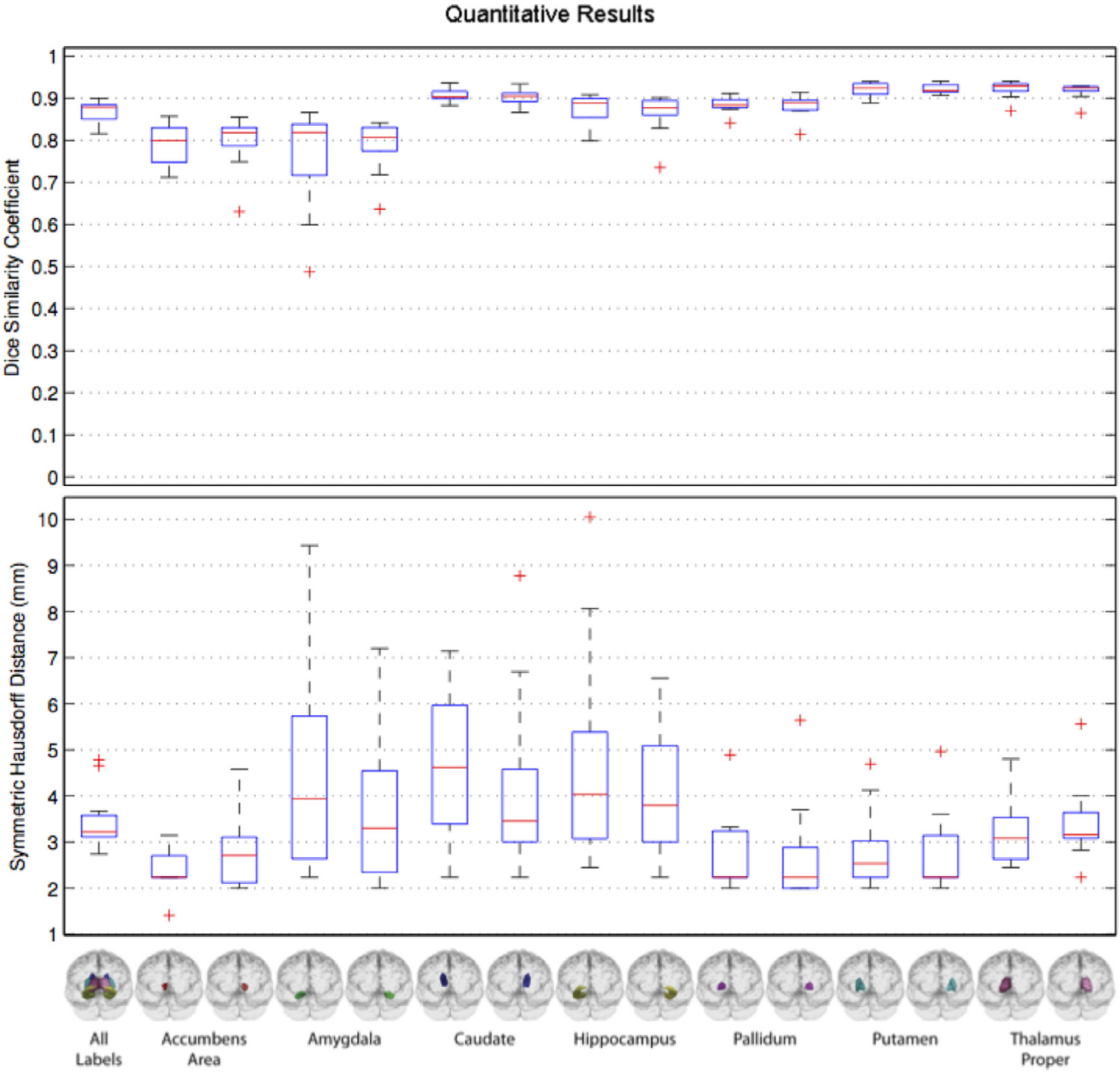

Fig. 12. Dice ratio and Hausdorff distance (in mm) achieved by our full method (MCfull) across different structures in the SATA Challenge dataset.

outperforms both the purely reconstruction-based methods and the purely classification-based methods, thus confirming the benefit of our transversal approach. Another interesting conclusion is that including the neighboring atlas patches into the dictionary leads to performance improvements, as shown by the comparison between LWV and NLWV and also confirmed by other studies (Rousseau et al., 2011). Finally, both the statistical estimation and the MRF-based regularization implemented by STEPS have proven beneficial for label fusion, as deduced when comparing the results of STEPS and LWV.

Appendix A. Implementation details

In this section, we describe the following details of our method, namely, (A.1) computation of a continuous label map from the overlapping estimations, (A.2) discretization based on confidence threshold, and (A.3) construction of the initial tentative dictionary.

A.1. Computation of continuous label map from the overlapping estimations

As result of matrix-completion based label fusion, we obtain a low-rank matrix with continuous target labels. Such continuous labels can be interpreted as confidence values, such that the higher the values above zero, the more likely to represent a foreground voxel, and the lower the values below zero, the more likely to represent a background voxel. Since we predict the label values for the entire target patch, we end up with multiple estimations (from the neighboring patches) for each target image point. We average the multiple estimations of each point in order to obtain a single value. The fusion process described here corresponds to the function $\boldsymbol{F} = \text{CombineOverlappingLabels}(F)$ in the algorithms of Table 1 and Table 2, where $\boldsymbol{F}$ is the continuous label map obtained by averaging the overlapping estimations contained in F.

A.2. Confidence-based discretization

At the end of each iteration, we compute the discrete label map $\boldsymbol{D}^{(s)}$ by assigning labels to the most confident voxels according to a confidence threshold τ, which is decreased at each iteration. This procedure is denoted by the function $\boldsymbol{D}^{(s)} = \text{Discretize}(\boldsymbol{F}, \tau)$ in the algorithm of Table 2 and is carried out as follows:

$$\boldsymbol{D}^{(s)}(x) = \begin{cases} 1 & \text{if } \boldsymbol{F}(x) \geq \tau \\ -1 & \text{if } \boldsymbol{F}(x) < -\tau \\ \perp & \text{Otherwise} \end{cases} \tag{A.1}$$

where $\boldsymbol{F}(x)$ denotes the confidence value at voxel x. Essentially, only the voxels with higher (in magnitude) confidence values above or below zero are assigned a label, whereas the voxels close to zero are left unassigned. As we decrease the confidence threshold, more ambiguous voxels are labeled. In the case of $\tau = 0$, all voxels are assigned a label regardless of their confidence values.

A.3. Construction of the initial tentative dictionary

Recall that, matrix-completion based label fusion in Section 3.1 uses a dictionary of atlas patches, denoted as (A, L), to label a specific target patch centered at position $x \in \Omega$. The sequential confidence-guided labeling framework in Section 3.2 further refines this initial tentative dictionary based on the label similarity. The dictionary building is denoted by the function $(A, L) = \text{BuildDictionary}(\boldsymbol{I}_k, \boldsymbol{L}_k, \vec{t}, x)$ in algorithms in Table 1 and Table 2. Both spatial proximity and appearance similarity to the target patch have been demonstrated to be good criteria to build the dictionary (Coupe et al., 2011; Rousseau et al., 2011). According to spatial proximity, we select the patches in the neighborhood of the target patch from *all* the atlases. That is, we build the dictionaries $A = [\vec{a}_1 \ldots \vec{a}_{\text{qm}}]$ and $L = [\vec{l}_1 \ldots \vec{l}_{\text{qm}}]$ from q patches of each of all m atlases in the neighborhood of the target patch. According to image similarity, we exclude the neighboring atlas patches whose appearance similarity with the target patch is below a certain image similarity threshold γ. Using the same criterion as in (Coupe et al., 2011), we only keep the atlas patches $\vec{a}_j$ satisfying the following equation:

$$\frac{2\mu_{\vec{t}}\mu_{\vec{a}_j}}{\mu_{\vec{t}}^2 + \mu_{\vec{a}_j}^2} \times \frac{2\sigma_{\vec{t}}\sigma_{\vec{a}_j}}{\sigma_{\vec{t}}^2 + \sigma_{\vec{a}_j}^2} > \gamma \tag{A.2}$$

where $\mu_{\vec{t}}$ and $\sigma_{\vec{t}}$ denote the mean and standard deviation of image patch $\vec{t}$ and $0 \leq \gamma \leq 1$ is the image similarity threshold.

Appendix B. Details of the classification-based method

As representative of the classification-based methods, we have implemented a label fusion variant closely related to our proposed method, i.e., using the logistic regression (**LogReg**) to learn the relationship between image appearance and anatomical labels of the atlas patches in the dictionary. The labels on each target patch, denoted as $\vec{f}$, are then computed as a mapping of its appearance vector, denoted as $\vec{t}$, by using the learned relationship, as follows:

$$\vec{f} = \text{logit}(V^\top \vec{t} + \vec{c}) \tag{B.1}$$

where logit(·) is the logistic function (a smoothed sign function), and V and $\vec{c}$ are the relationship matrix and bias vector, respectively. We learn the relationship between atlas appearance and labels using multi-task logistic regression[7] (Liu et al., 2009), where each label in the patch is encoded as an individual task. This corresponds to the following optimization problem:

$$\min_{V,\vec{c}} \text{C}_{\text{LL}}\left(L, V^\top \begin{bmatrix} A \\ \vec{1}^\top \end{bmatrix} + \vec{c}\right) + \alpha \|V\|_{\ell_1/\ell_2} \tag{B.2}$$

where $C_{\text{LL}}(\cdot)$ is the element-wise logistic loss between two matrices, and $\|V\|_{\ell_1/\ell_2}$ is the regularization enforcing sparsity across the rows of the matrix V and thus encouraging the sharing of features across different tasks (i.e., predictions of multiple labels in the target patch). The amount of regularization is controlled by the parameter α.

References

Aljabar, P., Heckemann, R.A., Hammers, A., Hajnal, J.V., Rueckert, D., 2009. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. NeuroImage 46, 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., 2013. Non-local statistical label fusion for multi-atlas segmentation. Med. Image Anal. 17, 194–208.

Cabral, R.S., De la Torre, F., Costeira, J.P., Bernardino, A., 2011. Matrix completion for multi-label image classification, NIPS.

Candès, E.J., Recht, B., 2009. Exact matrix completion via convex optimization. Found. Comput. Math. 9, 717–772.

Cardoso, M.J., 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, 671–684.

Collins, D.L., Pruessner, J.C., 2009. Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion. NeuroImage 52, 1355–1366.

Coupe, 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.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Royal Stat. Soc. Series B 39, 1–38.

Goldberg, A.B., Zhu, X., Recht, B., Xu, J.-M., Nowak, R.D., 2010. Transduction with Matrix Completion: Three Birds with One Stone. NIPS, pp. 757–765.

Hao, Y., Wang, T., Zhang, X., Duan, Y., Yu, C., Jiang, T., Fan, Y., 2013. Local Label Learning (LLL) for Subcortical Structure Segmentation: Application to Hippocampus Segmentation. Human Brain Mapping.

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, 115–126.

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.

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, 1000–1010.

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17, 825–841.

Jia, H., Yap, P.T., Shen, D., 2012. Iterative multi-atlas-based multi-image segmentation with tree-based registration. NeuroImage 59 (1), 422–430.

Kim, M., Wu, G., Li, W., Wang, L., Son, Y.-D., Cho, Z.-H., Shen, D., 2013. Automatic hippocampus segmentation of 7.0 Tesla MR images by combining multiple atlases and auto-context models. NeuroImage.

Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B.B., Chiang, M.C., Christensen, G.E., Collins, D.L., 2009. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. NeuroImage 46.

Li, G., Wang, L., Shi, F., Lyall, A.E., Lin, W., Gilmore, J.H., Shen, D., 2014. Mapping longitudinal development of local cortical gyrification in infants from birth to 2 years of age. J. Neurosci. 34, 4228–4238.

Li, K., Guo, L., Li, G., Nie, J., Faraco, C., Cui, G., Zhao, Q., Miller, L.S., Liu, T., 2010. Gyral folding pattern analysis via surface profiling. NeuroImage 52, 1202–1214.

Liu, J., Ji, S., Ye, J., 2009. SLEP: Sparse Learning with Efficient Projections. Arizona State University.

Nie, J., Li, G., Shen, D., 2013. Development of cortical anatomical properties from early childhood to early adulthood. NeuroImage 76, 216–224.

Nie, J., Li, G., Wang, L., Gilmore, J.H., Lin, W., Shen, D., 2011. A computational growth model for measuring dynamic cortical development in the first year of life. Cerebral Cortex 22, 2272–2284.

Rohlfing, R., Russakoff, D.B., Maurer, C.R. , 2004a. Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. IEEE Trans. Med. Imaging 23, 983–994.

Rohlfing, T., Brandt, R., Menzel, R., Maurer Jr, C.R., 2004b. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. NeuroImage 21, 1428–1442.

Rohlfing, T., Brandt, R., Menzel, R., Russakoff, D.B., Maurer, J., Calvin, R., 2005. Quo Vadis, Atlas-Based Segmentation?. The Handbook of Medical Image Analysis – Volume III: Registration Models. Kluwer Academic / Plenum Publishers.

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

Sanroma, G., Alquézar, R., Serratosa, F., 2012a. A new graph matching method for point-set correspondence using the EM algorithm and Softassign. Comput. Vision Image Understand. 116, 292–304.

Sanroma, G., Alquézar, R., Serratosa, F., Herrera, B., 2012b. Smooth point-set registration using neighboring constraints. Pattern Recognit. Lett. 33, 2029–2037.

Sanroma, G., Wu, G., Gao, Y., Shen, D., 2014a. Learning to rank atlases for multiple-atlas segmentation. to appear in IEEE Trans. Med. Imaging.

Sanroma, G., Wu, G., Thung, K.H., Guo, Y., Shen, D., 2014b. Novel multi-atlas segmentation by matrix completion. In: Wu, G., Zhang, D., Zhou, L. (Eds.), Machine Learning in Medical Imaging. Springer International Publishing, pp. 207–214.

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.

Shen, D., Davatzikos, C., 2002. HAMMER: hierarchical attribute matching mechanism for elastic registration. IEEE Transactions on Medical Imaging 21.

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

[7] We use the function mcLogisticR in the SLEP package from: http://www.public.asu.edu/~jye02/Software/SLEP

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

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

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

Wang, H., Das, S., Suh, J.W., Altinay, M., Pluta, J., Craige, C., Avants, B., Yushkevich, P. , 2011a. A learning-based wrapper method to correct systematic errors in automatic image segmentation: consistently improved performance in hippocampus, cortex and brain segmentation. NeuroImage 55, 968–985.

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

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.

Wu, G., Jia, H., Wang, Q., Shen, D., 2011. SharpMean: groupwise registration guided by sharp mean image and tree-based registration. NeuroImage 56 (4), 1968–1981.

Wu, M., Rosano, C., Lopez-Garcia, P., Carter, C.S., Aizenstein, H.J., 2007. Optimum template selection for atlas-based segmentation. NeuroImage 1612–1618.

Zhang, D., Guo, Q., Wu, G., Shen, D., 2012. Sparse patch-based label fusion for multi-atlas segmentation, multimodal brain image analysis, LNCS.

Zhang, D., Wu, G., Jia, H., Shen, D., 2011. Confidence-guided sequential label fusion for multi-atlas based segmentation, MICCAI.

Zhang, L., Wang, Q., Gao, Y., Wu, G., Shen, D., 2014. Learning of atlas forest hierarchy for automatic labeling of MR brain images, MLMI.

Zikic, D., Glocker, B., Criminisi, A., 2013. Atlas encoding by randomized forests for efficient label propagation, MICCAI.