NeuroImage 76 (2013) 11–23

Contents lists available at SciVerse ScienceDirect

NeuroImage

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

Segmentation of MR images via discriminative dictionary learning and sparse coding: Application to hippocampus labeling

Tong Tong [a,*], Robin Wolz [a], Pierrick Coupé [b], Joseph V. Hajnal [c], Daniel Rueckert [a], The Alzheimer's Disease Neuroimaging Initiative [1]

[a] Biomedical Image Analysis Group, Department of Computing, Imperial College London, 180 Queen's Gate, London, SW7 2AZ, UK
[b] LaBRI, CNRS UMR 5800, 351 cours de la Libération, F-33405 Talence, France
[c] Center for the Developing Brain, Division of Imaging Sciences and Biomedical Engineering, King's College London, St. Thomas Hospital, London, SE1 7EH, UK

ARTICLE INFO

Article history:
Accepted 25 February 2013
Available online 21 March 2013

Keywords:
Structural MR images
Patch-based segmentation
Discriminative dictionary learning
Sparse representation

ABSTRACT

We propose a novel method for the automatic segmentation of brain MRI images by using discriminative dictionary learning and sparse coding techniques. In the proposed method, dictionaries and classifiers are learned simultaneously from a set of brain atlases, which can then be used for the reconstruction and segmentation of an unseen target image. The proposed segmentation strategy is based on image reconstruction, which is in contrast to most existing atlas-based labeling approaches that rely on comparing image similarities between atlases and target images. In addition, we propose a Fixed Discriminative Dictionary Learning for Segmentation (F-DDLS) strategy, which can learn dictionaries offline and perform segmentations online, enabling a significant speed-up in the segmentation stage. The proposed method has been evaluated for the hippocampus segmentation of 80 healthy ICBM subjects and 202 ADNI images. The robustness of the proposed method, especially of our F-DDLS strategy, was validated by training and testing on different subject groups in the ADNI database. The influence of different parameters was studied and the performance of the proposed method was also compared with that of the nonlocal patch-based approach. The proposed method achieved a median Dice coefficient of 0.879 on 202 ADNI images and 0.890 on 80 ICBM subjects, which is competitive compared with state-of-the-art methods.

Introduction

The accurate and robust labeling of anatomical structures is an essential step in quantitative brain magnetic resonance imaging (MRI) analysis. Many clinical applications rely on the segmentation of MRI brain structures, which enables us to describe how brain anatomy changes during aging or disease progression. Since manual labeling by clinical experts is subject to inter and intra rater variability and is also a highly laborious task, an automated technique is desirable to enable a routine analysis of brain MRIs in clinical use. Despite the large number of existing techniques (Aljabar et al., 2009; Chupin et al., 2007; Collins and Pruessner, 2010; Coupé et al., 2011; Van der Lijn et al., 2008; Wang et al., 2011a; Wolz et al., 2010), it still remains a challenging task to develop fast and accurate automated segmentation methods due to the complexity of subcortical structures.

Many automated methods have been introduced to extract cortical and subcortical structures in the past decade. Among them, atlas-based methods have been shown to outperform other state-of-the-art algorithms (Babalola et al., 2009; Collins et al., 1995). In atlas-based label propagation, an atlas is matched to the target image using image registration. The segmentation of the target image is then achieved by warping the atlas label to the target image space. Segmentation errors produced by atlas-based methods can be classified into *random* errors and *systematic* errors (Aljabar et al., 2009; Wang et al., 2011a). *Random* errors, which may be caused by image noise or subject variation, can be reduced by using multiple atlases (Heckemann et al., 2006; Rohlfing et al., 2004) or by selecting the most similar atlases for a given unseen image (Aljabar et al., 2009; Artaechevarria et al., 2009; Barnes et al., 2008). *Systematic* errors occur consistently as the disagreement between manual and automatic segmentations exhibits a systematic pattern, which may be caused by consistent errors in the registration, partial volume effects or bias in the manual labeling of the atlases. For example, a manual segmentation protocol may follow a specific anatomical criterion to assign labels to different voxels. However, an automatic method may employ a slightly different criterion, which causes systematic labeling errors. Recent work has been proposed to reduce such errors such as intensity models (Lotjonen et al., 2010; Van der Lijn et al., 2008; Wolz et al., 2009) or a learning-based method (Wang et al., 2011a). Several label

* Corresponding author.
E-mail address: t.tong11@imperial.ac.uk (T. Tong).

[1] Data used in the preparation of this article were obtained from the ADNI database (www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: www.loni.ucla.edu/ADNI/Collaboration/ADNI_Authorship_list.pdf.

1053-8119/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.02.069

fusion techniques have also been used to improve segmentation accuracy of the multi-atlas segmentation method such as majority voting (Aljabar et al., 2009; Collins and Pruessner, 2010; Kittler, 1998), STAPLE (Warfield et al., 2004) and local weighted label fusion (Artaechevarria et al., 2009; Khan et al., 2011; Sabuncu et al., 2010; Wang et al., 2011b, 2011c). However, multi-atlas segmentation requires pairwise, accurate registrations between atlas and target, which can result in a significant computational burden.

Recently, nonlocal patch-based segmentation techniques have been proposed (Coupé et al., 2011, 2012; Rousseau et al., 2011) to avoid the need of accurate non-rigid registration in order to gain computational efficiency. Instead of fusing propagated label maps as in multi-atlas segmentation, this method obtains a label for every voxel by using similar image patches from coarsely aligned atlases. First, image patches are extracted in a predefined neighborhood around a particular voxel and across the training atlases. Then, weights are given to these patches according to the similarity between the target patch and the extracted atlas patches. The final label of the target voxel is estimated by fusing the labels of the central voxels in the template patch library. Such a technique allows one-to-many correspondences to select the most similar patches for label fusion, and a validation on hippocampus segmentation (Coupé et al., 2011, 2012) demonstrates a high accuracy of this approach.

Although the local weighting label fusion strategy (Artaechevarria et al., 2009; Khan et al., 2011; Sabuncu et al., 2010; Wang et al., 2011b, 2011c) or the nonlocal patch-based technique (Coupé et al., 2011, 2012; Eskildsen et al., 2011; Rousseau et al., 2011) can produce accurate segmentation results, these methods are based on the similarity of image patches extracted from each atlas. However, image similarities over small image patches may not be an optimal estimator (Wang and Yushkevich, 2012). In this paper, we propose a novel segmentation method based on image patch reconstruction. The proposed approach uses discriminative dictionary learning methods (Jiang et al., 2011; Yang et al., 2011a; Zhang and Li, 2010) and sparse coding techniques (Wright et al., 2009). These methods have been successfully applied to different problems in face recognition (Jiang et al., 2011; Wright et al., 2009; Yang et al., 2011a; Zhang and Li, 2010). To the best of our knowledge, these methods have never been used in subcortical brain segmentation. The proposed method learns discriminative appearance dictionaries and is different from the recent work in Zhang et al. (2012), which learns shape dictionaries for liver segmentations. In the proposed method, we abandon the conventional idea to compare the similarity between patches in a neighborhood. Instead, a dictionary and a linear classifier are learned from the template patch library simultaneously for every voxel in the target image. The surrounding patch of the target voxel can be reconstructed by the corresponding dictionary and the label of the target voxel will be estimated by the corresponding classifier. Moreover, a new strategy has been proposed to implement the method in an efficient way by learning dictionaries offline and performing segmentation online.

In the remainder, we will first introduce the methodology of discriminative dictionary learning and how we apply it to the segmentation of brain MR images. The proposed method was evaluated on hippocampus segmentations on 202 ADNI images (Mueller et al., 2005) and 80 healthy ICBM subjects (Mazziotta et al., 1995). We studied the influence of different parameters and compared the performance of the proposed methods with that of the nonlocal patch-based technique. The performance of different methods has been compared on different subject groups of the ADNI dataset. Finally, we discuss the strengths and weaknesses of the proposed method and conclude the paper.

Table 1
Demographic information describing 202 ADNI images used in this study.

	Number	Age	MMSE
Normal	68	76.31 ± 5.20 [62–88]	29.18 ± 0.88 [26–30]
SMCI	49	74.96 ± 7.28 [60–89]	27.55 ± 1.67 [24–30]
PMCI	44	75.38 ± 6.92 [60–88]	26.80 ± 1.69 [24–30]
AD	41	76.08 ± 7.23 [56–87]	23.12 ± 1.79 [20–26]

Materials and methods

Datasets

Two different datasets were used for hippocampus segmentations. Images obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI) (Mueller et al., 2005) and images obtained from the International Consortium for Brain Mapping (ICBM) database (Mazziotta et al., 1995) were used to evaluate the proposed approach.

In the ADNI study, brain MR images are acquired at regular intervals from approximately 200 cognitively normal older individuals, 400 people with Mild Cognitive Impairment (MCI), and 200 people with early AD. A more detailed description of the ADNI study is given in A. The subgroup we used consists of 202 subjects obtained from different scanners (68 normal subjects, 93 subjects with MCI and 41 patients with AD). An overview of these 202 subjects is shown in Table 1. These 202 images were selected because their reference segmentations are available through ADNI. The selected subgroup is representative of the whole ADNI dataset as no significant difference was observed on age and MMSE scores between the selected group and the whole dataset on Student's t-test ($p > 0.1$). A commercially available high dimensional brain mapping tool (Medtronic Surgical Navigation Technologies, Louisville, CO) was used to carry out semi-automated hippocampal volumetry for defining these reference segmentations. These label maps were inspected and if necessary manually corrected by qualified reviewers (Hsu et al., 2002).

For a direct comparison with the previously published patch-based method, the proposed method was also evaluated on a subset of the ICBM dataset, which consists of 80 healthy subjects (Mazziotta et al., 1995). The T1-weighted data were acquired at the Montreal Neurological Institute on a Philips Gyroscan 1.5 T scanner with 3D spoiled gradient-echo acquisition with TR = 17 ms, TE = 10 ms, flip angle = 30°, and a resolution of 1 mm^3 voxels. The 80 subjects consist of 39 males and 41females of similar ages (mean age: 25.09 ± 4.9 years). The MR images were manually segmented by an expert directly in stereotaxic space using the protocol described in Pruessner et al. (2000). The resulting segmentations obtained an intraclass reliability coefficient (ICC) of 0.900 for inter-rater reliability (4 raters) and 0.925 for intra-rater reliability (5 repeats).

Overview of the method

The basic assumption of non-local means patch-based segmentation is that the central voxels of similar patches are considered to belong to the same structure (Coupé et al., 2011). This method assigns higher weights to similar patches and smaller weights to dissimilar patches. As a result, similar patches from each training atlas contribute more to the final label estimation. The assumption of our method is that the target patch which will be labeled can be represented by a few template patches from the same structure in a low-dimensional manifold or by a few representative atoms from a learned dictionary. After the coding of the target patch, the target voxel is labeled based on the coding coefficients and the dictionary. Therefore, there are two phases for labeling a voxel in the proposed method: coding and classification. In the proposed method, a different dictionary is learned for labeling each different target voxel, which means that the learned dictionaries are voxel based. Based on different types of dictionaries used for coding, we divided our proposed method into two groups: sparse representation classification (SRC) (Tong et al., 2012) and Discriminative Dictionary Learning for Segmentation (DDLS).

For SRC, the whole template patch library is directly defined as the dictionary for sparse coding. In this predefined dictionary, each atom is a patch extracted from an atlas image and we then know the corresponding labels of all the atoms. After the target patch is coded by this predefined dictionary, the labeling of the target voxel is done by assessing which group of patches provides the minimal reconstruction error.

However, the direct way of using all the training patches as the dictionary may result in a huge size for the dictionary, increasing the coding complexity. In addition, this predefined dictionary may not fully exploit the discriminative information in the training patch library. In this paper, we also extend the proposed SRC method by learning discriminative dictionaries for segmentation. In this DDLS method, a small-sized dictionary and a linear classifier are learned from the template training patch library, which will provide reconstructive and discriminative information for MR brain segmentation work. Fig. 1 demonstrates the proposed segmentation process for one target voxel and compares the major differences with non-local means patch-based methods.

In the proposed SRC and DDLS approaches, the coding procedure by dictionaries replaces the patch similarity weighting in the patch-based segmentation strategy. However, each subject is segmented by learning specific dictionaries from the atlas database, which will be computationally expensive. The ultimate goal of our work is to learn fixed dictionaries and classifiers offline. Then, segmentations can be performed online by using the fixed dictionaries and classifiers. In order to do this, we also proposed a Fixed Discriminative Dictionary Learning for Segmentation (F-DDLS) strategy, which could enable a significant speed-up in the segmentation phase. This F-DDLS strategy has been evaluated on the ICBM dataset and different subject groups of the ADNI dataset.

Sparse representation classification

The SRC method uses a template patch library as a predefined dictionary. First, the extraction of a patch library from atlases will be introduced. Then, this predefined dictionary will be used for the reconstruction of a target patch and the sparse coding process will be introduced. Finally, the sparse representation of the target patch and the labels of center voxels of template patches will be used for estimating the label of the target voxel.

Construction of patch library: First, atlas selection is carried out for every target subject based on the sum of squared intensity differences (SSD) in a template space as done in Coupé et al. (2011). Then for labeling a target voxel in the target image, the surrounding patch (illustrated by the red box overlaid on the target image in Fig. 1) is extracted and denoted as the target patch p_t in this paper. A search volume V_i (illustrated by the blue box overlaid on the atlas images in Fig. 1) is defined in each atlas image. All template patches in the search volume across a set of similar atlases are extracted to form a patch library. The number of patches in the patch library is proportional to the search volume size

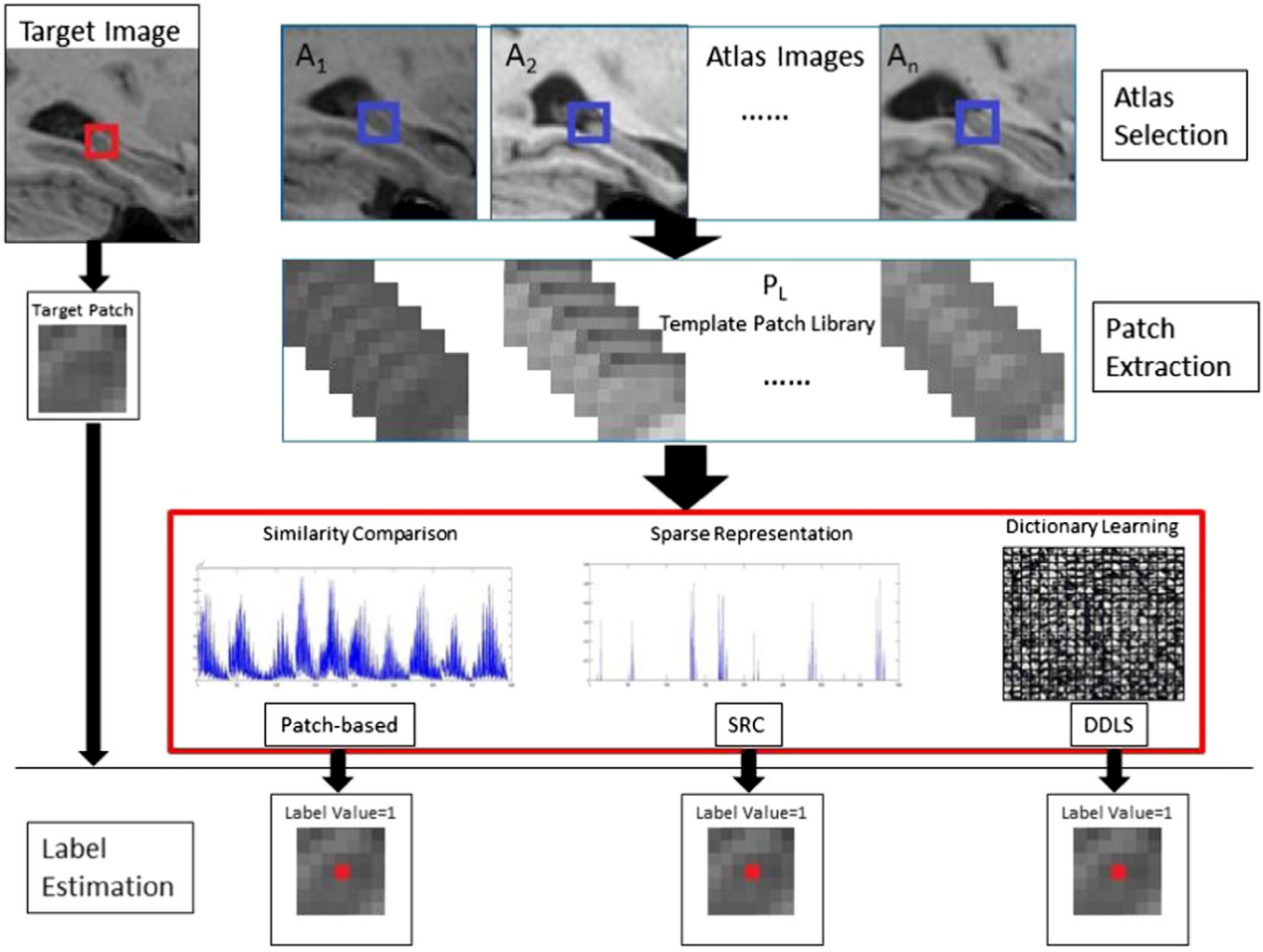

Fig. 1. Flow chart of labeling one target voxel by three different methods: Patch-based Labeling, sparse representation classification (SRC) and Discriminative Dictionary Learning for Segmentation (DDLS). The red box in the target image represents the target patch. The blue boxes in atlas images represent the search volume area for extracting template patches.

and typically contains thousands of patches. Each patch in the library is a volume. We denote each patch as a column vector and group all the patches together as a matrix P_L. Suppose that the patch library contains n patches, then the patch library can be represented as $P_L = [p_1,p_2,\cdots,p_n] \in R^{m \times n}$.

Inspired by work in face recognition (Wright et al., 2009), we propose to use a sparse representation classification strategy for patch selection and weighting. In the SRC method, the patch library is directly considered as a dictionary, so the target patch p_t will approximately lie in the subspace spanned by the training patches in the library P_L:

$$p_t = a_1 p_1 + a_2 p_2 + \cdots + a_n p_n. \tag{1}$$

Since the SRC method imposes a constraint that the representation is sparse, most of the coefficients a_i will be zero. Let $a = [a_1,a_2,\cdots,a_n] \in R^n$, then the sparse solution can be obtained by solving the following equation:

$$\hat{a} = \min_{a} \|a\|_0 \ \text{ subject to } \ \|p_t - P_L a\|_2^2 \le \varepsilon \tag{2}$$

where the l_0-norm denotes the number of nonzero coefficients, which is the sparse constraint of this equation. The linear system $p_t = P_L a$ is underdetermined since $n > m$, so this equation does not have a unique solution. It is difficult to approximate the sparsest solution of an underdetermined system of linear equations because the problem is NP-hard. In general, if the solution of Eq. (2) is sparse enough, then it can be shown to be equivalent to the solution of the following l_1-minimization problem (Wright et al., 2009):

$$\hat{a} = \min_{a} \|a\|_1 \ \text{ subject to } \ \|p_t - P_L a\|_2^2 \le \varepsilon. \tag{3}$$

Eq. (3) can be solved efficiently by several sparse coding methods (Yang et al., 2010). In Wright et al. (2009), Eq. (3) was solved using the Lasso method (Tibshirani, 1996). The L_1 Lasso is a relaxed version of Eq. (3). However, if the number of predictors (n) is much higher than the number of observations (k), the Elastic Net (EN) approach always outperforms the Lasso method for achieving a satisfactory variable selection (Zou and Hastie, 2005). Considering that the number of patches in the library is much higher than the number of patches selected for representation, our case belongs to this 'large n small k' problem. To achieve robust sparse representations, EN (Zou and Hastie, 2005) has been used for obtaining the sparse coding coefficients:

$$\hat{a} = \min_{a} \frac{1}{2}\|p_t - P_L a\|_2^2 + \lambda_1 \|a\|_1 + \frac{\lambda_2}{2}\|a\|_2^2. \tag{4}$$

Eq. (4) adds a coefficient magnitude penalty to the objective function in Eq. (3), which is a convex combination of L_1 lasso and L_2 ridge penalties. EN encourages a grouping effect while keeping a similar sparsity of representation (Zou and Hastie, 2005). This grouping effect, which selects groups of highly correlated variables, is helpful for the final classification and could thus improve the segmentation performance.

After we obtain the sparse solution, the labeling of the target voxel is based on the coding coefficients $\hat{a}$ and the selected patches for representation. The main idea is that the sparse nonzero coefficients should concentrate on the training patches with the same class label as the target patch. This means that the training patches from the correct class will yield the minimal reconstruction error when the coding coefficients are computed using training patches from all classes. There are two key points in our assumption. First, the coding coefficients are very sparse. In fact, both the training patches from the correct class and the wrong class can represent the target patch very well if enough training patches from each class are given. However, when the sparsity is imposed on the coding coefficients, each class can only use a few patches to represent the target patch. In this case, the training patches from the correct class are likely to represent the target patch with less error. Second, the coding coefficients are computed using all classes, which also helps for classification. This is because the training patches from each class will compete to represent the target patch in the same process. These two points were verified in Yang et al. (2011b). Therefore, the labeling of the target voxel is achieved by comparing which class of training patches gives the minimal reconstruction error. The residual (reconstruction error) with the sparse coefficients $\hat{a}^j$ associated to each structure/class j is described as:

$$r_j(p_t) = \left\| p_t - P_L^j \hat{a}^j \right\|. \tag{5}$$

The label value v for the center voxel of target patch p_t is assigned as the class with the minimum residual over all classes:

$$v = \underset{j}{argmin} \left(r_j(p_t) \right) \ j = 1, \ldots C. \tag{6}$$

In our case, the patches associated with non-zero coefficients are divided into two groups ($C = 2$): patches belonging to the hippocampus and patches belonging to the background. If the patches belonging to the hippocampus can represent the target patch p_t with a smaller reconstruction error, then the target voxel is labeled as hippocampus, and vice versa.

Discriminative dictionary learning

In the above SRC scheme, a large number of training patches are directly used as the predefined dictionary, which will increase the computational burden on the sparse coding process. Also, this predefined dictionary may not fully exploit the discriminative information in the training patch library. These drawbacks may be overcome by learning a small-sized task-specific dictionary. Several methods have been proposed for learning a small-sized dictionary (Jiang et al., 2011; Mairal et al., 2008, 2009b; Yang et al., 2011a; Zhang and Li, 2010) that has good reconstructive power and discriminative ability. In particular, the method proposed in Zhang and Li (2010) incorporated the classification error into the objective function of the K-SVD algorithm, which allows to learn the dictionary and the classifier by the same optimization procedure simultaneously. In this paper, we used a similar idea as described in Zhang and Li (2010) for our segmentation purpose. Let $P_L = [p_1,p_2,\cdots,p_n] \in R^{m \times n}$ denote the training patch library, containing n patches. A reconstructive dictionary with K atoms can be learned from the input patch library P_L by solving the following problem:

$$\langle D, \alpha \rangle = \underset{D,\alpha}{argmin} \|P_L - D\alpha\|_2^2 \ \text{ subject to } \ \|\alpha\|_0 \le T \tag{7}$$

where $D = [d_1,d_2,\cdots,d_K] \in R^{m \times K}$ is the learned dictionary. $\alpha \in R^{n \times K}$ is the sparse coding coefficient matrix of the input patch library, and T is a sparsity constraint parameter. In Eq. (7), the objective function includes the reconstruction error term and the sparsity constraint term without considering the discriminative power. Thus, the learned dictionary is not suitable for our classification task. To address this problem, a linear classifier $f(\alpha,W) = W\alpha$ as in Zhang and Li (2010) was added to the objective function for learning dictionaries with both reconstructive and discriminative power. The objective function can then be defined as follows:

$$\begin{aligned} \langle D, W, \alpha \rangle = \underset{D,W,\alpha}{argmin} & \|P_L - D\alpha\|_2^2 + \beta_1 \|H - W\alpha\|_2^2 \\ \text{subject to } & \|\alpha\|_0 \le T \end{aligned} \tag{8}$$

where the classification error term $\|H - W\alpha\|_2^2$ is added to Eq. (7). H represents the labels of the central voxels of the patches in the library P_L. Each

column of H is a label vector corresponding to a template patch. Each label vector is defined as $h_i = [0,0\ldots1\ldots0,0]$, where the non-zero entry position indicates the label of the center voxel of the corresponding patch. W denotes the linear classifier parameters and β_1 controls the trade-off between the reconstruction error term and the classification error term.

In Jiang et al. (2011) and Zhang and Li (2010), the problem of Eq. (8) was solved by using the K-SVD algorithm since only one dictionary was required for the face recognition task under study. However, due to the high anatomical variability across subjects' brain scans, it is difficult to achieve good segmentation performance by just learning one dictionary and a single global classifier. Therefore, in our work, a dictionary and a corresponding classifier are learned for every target voxel. In this case, there will be thousands of dictionaries to be learned for every target subject and it will be very computationally expensive if the K-SVD algorithm is used for solving Eq. (8). Here, we use an online dictionary learning algorithm as proposed in Mairal et al. (2009a), which has faster performance and generates better dictionaries than classical batch learning algorithms. Appendix B shows how Eq. (8) is solved by using the online dictionary learning algorithm.

After Eq. (8) is solved, a dictionary $\hat{D}_t$ and a classifier $\hat{W}$ are learned for every target voxel. Fig. 2 shows an example of the learned dictionary $\hat{D}_t$. For labeling the target voxel, the surrounding patch p_t is first extracted. Then, the sparse representation $\hat{\alpha}_t$ of the target patch p_t is computed by solving the following problem:

$$\hat{\alpha}_t = \underset{\alpha_t}{argmin} \left\| p_t - \hat{D}_t \alpha_t \right\|_2^2 + \beta_2 \|\alpha_t\|_1. \tag{9}$$

Eq. (9) is a relaxed version of Eq. (3). Finally, we can estimate the label value v of the target voxel by using the linear predictive classifier:

$$\begin{cases} h_t = \hat{W}_t \hat{\alpha}_t \\ v = \underset{j}{argmax}\ h_t(j) \end{cases} \tag{10}$$

where h_t is the class label vector for the target voxel. The label value v of the target voxel is decided by the index of the largest element in label vector h_t. Ideally, h_t will be {0,0, ···,1, ···,0,0} with only one non-zero element, indicating the label of the class. In our binary segmentation, there are only two elements in label vector h_t, the values of which indicate the probability belonging to hippocampus and the probability belonging to the background respectively.

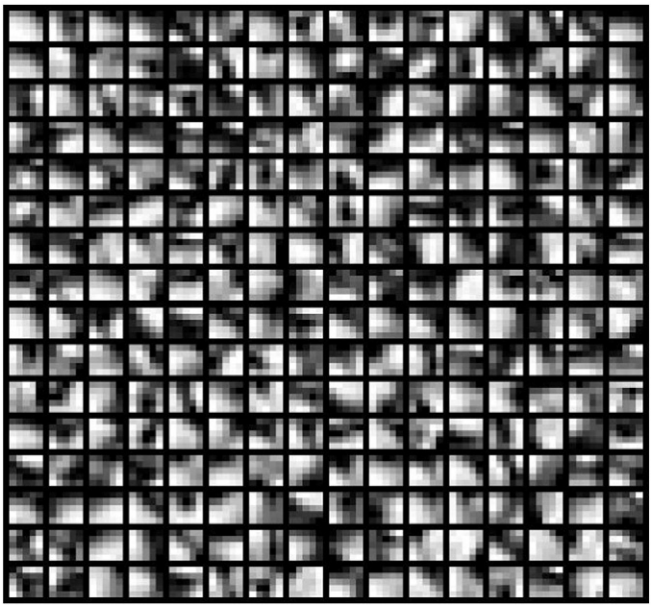

Fig. 2. An example of a learned dictionary. The dictionary has 16 × 16 atoms of size 5^3. This figure shows a slice of this dictionary.

Experiments and results

The proposed methods were applied to 202 images from the ADNI database and 80 images from the ICBM database. The ADNI images were preprocessed by the ADNI pipline described in Jack et al. (2008) and the ICBM images were preprocessed as described in Coupé et al. (2011, 2012). All images were linearly registered to the MNI152 template space by using affine registrations. Image intensities were then normalized by using the method proposed in Nyul and Udupa (2000). After that, intensities were rescaled to the interval [0 100]. Finally, a leave-one-out procedure was used in our validation and the most similar subjects were selected by comparing the squared intensity differences (SSD) in the MNI152 template space as described in the "Sparse representation classification" section.

For the ADNI images, all segmentations were performed in the native image space because the reference segmentations of the hippocampus are defined in native space. Transforming the labels and MR images into template space would decrease label accuracy due to interpolation artifacts of the target reference segmentations. After atlas selection in the MNI152 template space, we affinely transformed the selected atlases and labels to the native space of the target image to perform the segmentation in the target coordinate system.

For the ICBM images, all the segmentations were performed in the MNI152 template space as the labels are defined in the template space. The influence of different parameters was studied on segmenting the hippocampus on the 202 ADNI images. After the optimal parameters were estimated, both datasets were used to compare the performance of different methods.

For learning dictionaries, the parameter β_1 in Eq. (B.2) was set to 1 and β_2 was set to 0.15 for all experiments. β_2 was determined via cross validation according to the parameter settings described in Mairal et al. (2012). During parameter optimization, when a certain parameter was optimized, the other parameters were set to fixed values. Since neighborhood voxels share most of the template patches and will have very similar dictionaries and classifiers, we used a sampling strategy to train the dictionaries in order to achieve a better performance. Dictionaries are trained for every n ($n > 1$) voxels rather than every voxel. In theory, this strategy will achieve an approximate n^3 speedup for the training process. In order to label target voxels without a corresponding dictionary, we use neighbor dictionaries to perform sparse coding for its target patch. Since neighborhood voxels share most of their template patches and will have very similar dictionaries and classifiers, this sampling strategy will not result in a dramatic degradation of the segmentation accuracy. In our 3D segmentation work, we used 6 nearest neighbor dictionaries for sparse coding for all experiments. By using the same classification strategy, we could obtain 6 class label vectors for the target voxel. The final label value is estimated by using the average of these label vectors. Finally, all the experiments were evaluated by computing the Dice coefficient between the reference segmentations and the automated segmentations.

Influence of parameters for dictionary training

First, experiments were carried out to study the influence of the dictionary size K (the number of atoms in each dictionary) on segmentation accuracy. 10 atlases from the ADNI dataset were selected in a leave-one-out procedure for each target image. We found that the larger the size of dictionaries was, the higher the achieved overlap value was. However, the improvement using $K > 256$ over $K > 256$ is not significant and more time is required for learning a larger size of dictionaries. Considering the trade-off between computational time

and segmentation accuracy, we chose $K > 256$ for the following experiments according to the results shown in Fig. 3.

Since a sampling strategy was used for learning dictionaries, the influence of the sampling step size l on segmentation accuracy was also studied. The sampling step size l means that dictionaries are trained for every lth voxel. As expected, the smaller sampling step size we used for learning dictionaries, the more dictionaries we could get and the higher median Dice index could be achieved. However, we could gain more computational efficiency by learning fewer dictionaries. Based on the results shown in Fig. 4, a sampling step size of 3 (every third voxels for learning dictionaries) is a good choice for balancing the speed and accuracy of the proposed method.

Influence of patch size and neighborhood size

The influence of patch size and neighborhood size was also studied on the ADNI dataset. The patch size is related to the local geometry and the neighborhood size reflects the anatomical variability (Coupé et al., 2011). The Dice coefficient distributions over varying patch and neighborhood sizes are presented in Figs. 5 and 6. The best median Dice coefficient was obtained with a patch size of $5 \times 5 \times 5$ and a neighborhood size of $7 \times 7 \times 7$. Therefore, we used these parameter settings for all other experiments.

Influence of the number of training atlases

We also studied the impact of the number of training atlases on the performance of the proposed method. The results for varying numbers of training atlases out of the 202 ADNI images are shown in Fig. 7. As can be seen, increasing the number of training atlases provides higher median Dice overlap values. The best median Dice coefficient is 0.879 by selecting 25 atlases, which is a small improvement in comparison with a median Dice value of 0.872 by using 10 atlases. As can be seen in Fig. 7, the median Dice coefficient stabilizes around 10 atlases, which is similar to the trends reported in Coupé et al. (2011) and Rousseau et al. (2011). Therefore, we selected 10 atlases in our experiments as this produces comparable results with a significantly lower computational burden comparing to selecting 25 atlases, making the method more efficient and attractive.

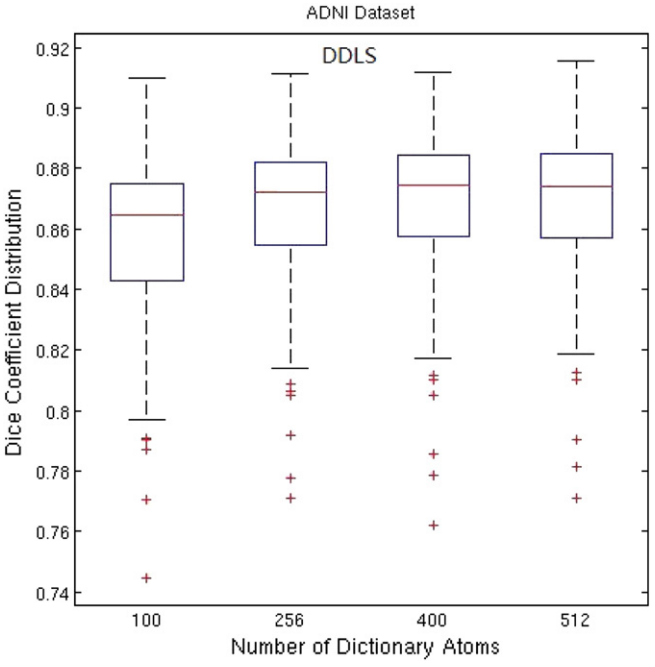

Fig. 3. Effect of dictionary size on segmentation accuracy. The results were obtained by using a patch size of $5 \times 5 \times 5$ voxels and a search volume of $7 \times 7 \times 7$ voxels, extracted from the 10 most similar atlases. The sampling step size was set to 3 for learning dictionaries.

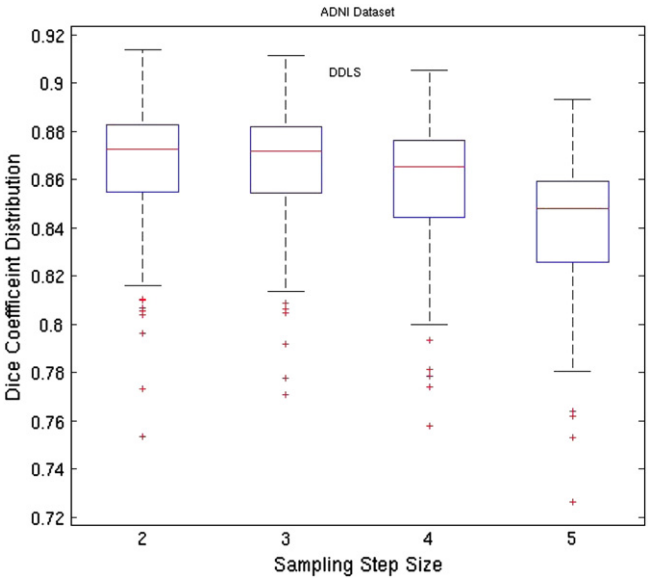

Fig. 4. Effect of sampling step size on segmentation accuracy. The results were obtained by using a patch size of $5 \times 5 \times 5$ voxels and a search volume of $7 \times 7 \times 7$ voxels, extracted from the 10 most similar atlases.

Fixed Discriminative Dictionary Learning for Segmentation

The ultimate goal of our work is to learn fixed dictionaries offline. Then the learned dictionaries can be used to efficiently perform segmentation online. In order to do this, we randomly selected a subgroup of the whole dataset as the training atlases. Then discriminative dictionaries were trained from these randomly selected training atlases. Finally, the same testing procedure as described in the "Discriminative dictionary learning" section" was performed to segment the remaining testing subjects. The F-DDLS strategy was evaluated on the 80 healthy ICBM subjects. 40 images were randomly selected for training dictionaries and classifiers. The remaining 40 atlases were used for testing. The experiment was repeated 10 times. The results are presented in Fig. 8. The average median Dice coefficient is 0.887. These results indicate

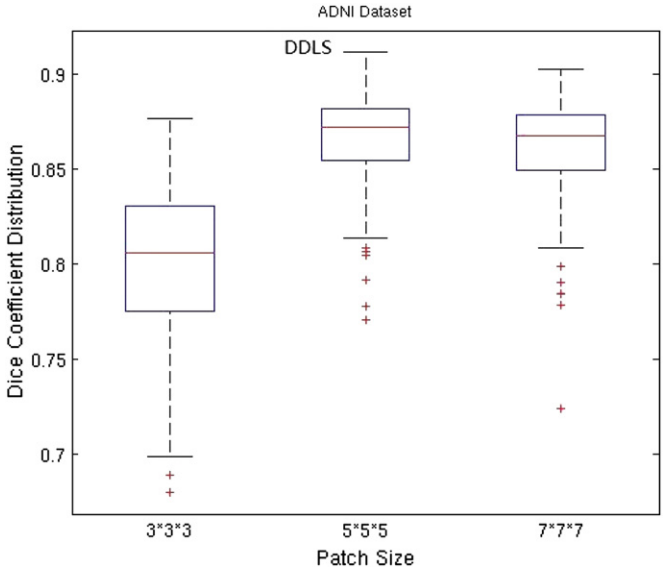

Fig. 5. Effect of patch size on segmentation accuracy. The results were obtained by using a search volume of $7 \times 7 \times 7$ voxels, extracted from the 10 most similar atlases.

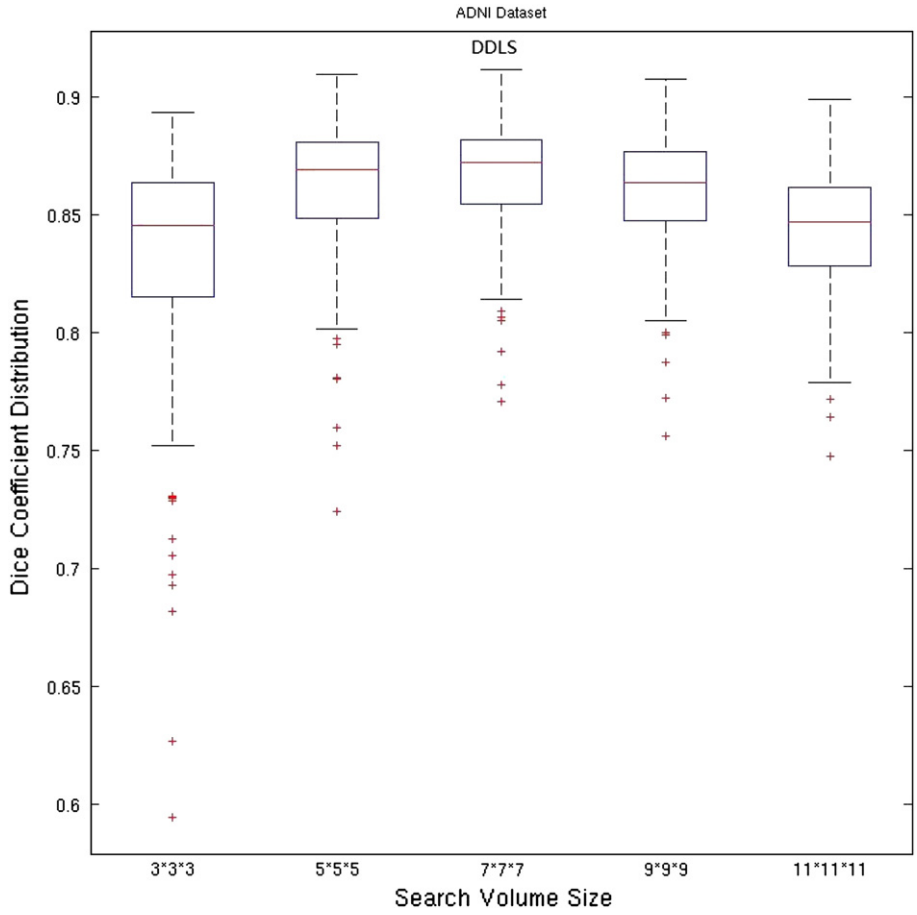

Fig. 6. Effect of neighborhood size on segmentation accuracy. The results were obtained by using a patch size of $5 \times 5 \times 5$ voxels, extracted from the 10 most similar atlases.

that the proposed F-DDLS strategy can be performed in a computationally more efficient way, while yielding comparable segmentation results. In particular, each unseen subject can be segmented in less than 1 min by using the proposed F-DDLS strategy.

Discriminative Dictionary Learning for Segmentation using Fixed Training Dataset

For the ICBM dataset, all images were transformed to MNI template space and the reference segmentations were then carried out in this space. Since label images and MRIs are in the same space for the whole dataset, fixed dictionaries can directly be learned in this common space. For the ADNI dataset, the segmentations were performed in the target image space because the reference segmentation is defined in this space. We used a work-around to simulate a common space for the whole dataset: after randomly selecting a fixed training subgroup, dictionaries were learned after warping the training images to the testing target image space. Although the learned dictionaries were not fixed in different spaces, they were learned from the fixed training dataset. Therefore, the dictionaries can be regarded as identical dictionaries transformed to different coordinate spaces.

We compared the results by using different numbers of training atlases in the fixed subgroup. The results are presented in Fig. 9. By using the simulated fixed discriminative dictionary learning strategy, the median Dice coefficient is 0.864 when using 30 subjects for training and the remaining 172 subjects for testing. Although this is slightly lower than the median Dice coefficient (0.879) by selecting 25 atlases in a leave-one-out procedure, it also demonstrates the effectiveness of the proposed method.

We also tested the proposed simulated F-DDLS method on different groups of data from ADNI. 30 healthy subjects were randomly selected for training fixed dictionaries. Segmentations were then performed on the 41 AD subjects by using the learned dictionaries from healthy subjects. The performance of different methods on the 41 AD subjects is presented in Fig. 10. It can be seen that DDLS achieved the highest mean Dice coefficient of 0.866 for the segmentations of the 41 AD subjects, which is much higher than a mean Dice coefficient of 0.826 by using the nonlocal patch-based method. The mean Dice coefficient is 0.860 for the 41 AD subjects by using the simulated F-DDLS method, which shows that the proposed method is able to generalize from one type of data to morphologically different datasets.

Comparison with other methods

The proposed DDLS and SRC were compared with the nonlocal patch-based technique proposed in Coupé et al. (2011). For the proposed DDLS method, we used a sampling step size of 3, a patch size of $5 \times 5 \times 5$ and a neighborhood size of $7 \times 7 \times 7$ as suggested in the "Influence of patch size and neighborhood size" section. For SRC, 80 patches were selected for representation and classification. λ_1 and λ_2 were set to 0.15 for obtaining the sparse representation to solve Eq. (4). These two parameters were determined via cross validation following the parameter settings described in Mairal et al. (2012). For a fair comparison, the nonlocal patch-based method was carried out in the same settings (a patch size of $7 \times 7 \times 7$ voxels and a search volume of $9 \times 9 \times 9$ voxels) as described in Coupé et al. (2011). The same patch preselection process as described in Coupé et al. (2011) was performed for both patch-based method and SRC method in order to reduce computational time.

For a direct and fair comparison with the nonlocal patch-based method, hippocampus segmentations were performed on the 80 healthy ICBM subjects in the MNI152 template space (Table 2). 20 atlases were selected

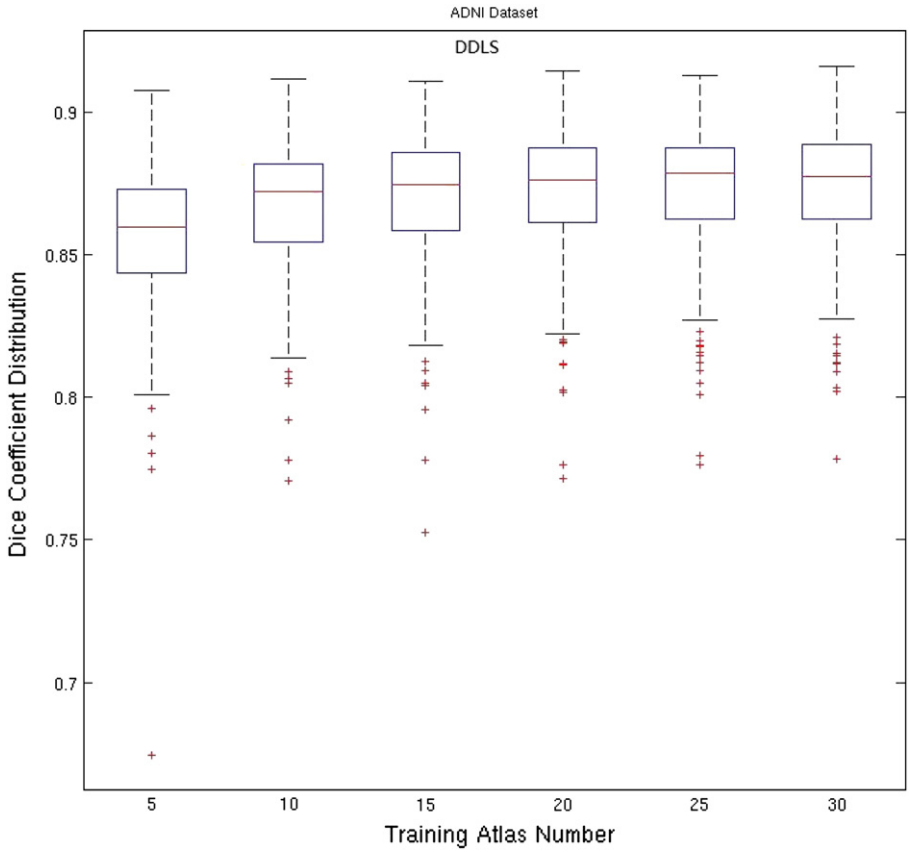

Fig. 7. Effect of the number of training atlases on segmentation accuracy. The results were obtained by using a patch size of 5 × 5 × 5 voxels and a search volume of 7 × 7 × 7 voxels.

in a leave-one-out procedure for each target image as suggested in Coupé et al. (2011). The patch-based method obtained a median Dice value of 0.882, the proposed SRC approach obtained 0.888, and the proposed DDLS obtained 0.890. Although the SRC method produces similar results as those produced by DDLS, the latter can be implemented with a much faster speed than the SRC method (as discussed in the next section). The DDLS method obtained significantly better results than the patch-based method with a p-value $\ll 0.001$ using Student's two-tailed paired *t*-test.

Hippocampus segmentations on the 202 ADNI images were performed in the native image space. 10 atlases were selected in a leave-one-out procedure for each target image. Table 3 presents the

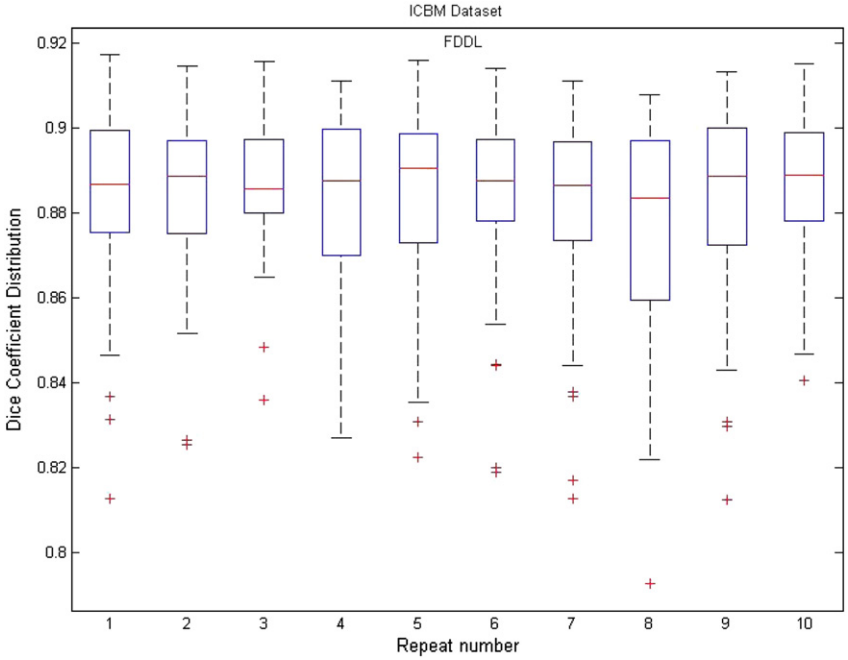

Fig. 8. The performance of Fixed Discriminative Dictionary Learning on 80 ICBM subjects. The results were obtained by using a patch size of 5 × 5 × 5 voxels and a search volume of 7 × 7 × 7 voxels. The experiment was repeated 10 times. The average median Dice coefficient is 0.887.

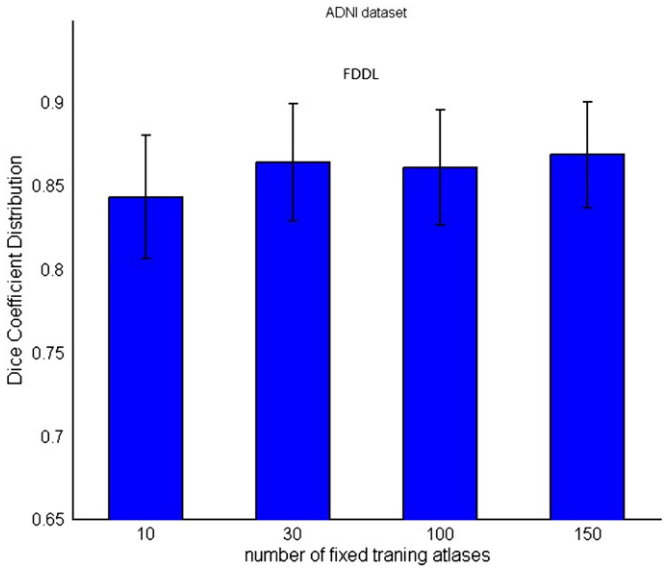

Fig. 9. Effect of the number of training atlases on the performance of Fixed Discriminative Dictionary Learning. The results were obtained by using a patch size of $5 \times 5 \times 5$ voxels and a search volume of $7 \times 7 \times 7$ voxels. The median Dice coefficient is 0.864 when using 30 subjects for offline training and the remaining 172 subjects for online testing.

median Dice coefficients of these three approaches. The median Dice coefficient is 0.872 by using the proposed DDLS and 0.871 by using SRC, both of which is higher than the median Dice value of 0.844 by using the nonlocal patch-based method. Fig. 11 provides a visual comparison of the segmentation results by using these three methods. We also compared the performances of these three methods on four different groups of subjects. Fig. 12 shows the mean Dice coefficients for the segmentations of 68 control subjects, 49 stable MCI subjects, 44 progressive MCI subjects and 41 AD patients. As revealed in Fig. 12, the segmentation accuracy decreases with disease progression, which indicates that smaller hippocampi, due to atrophy, are more challenging for automated segmentation approaches.

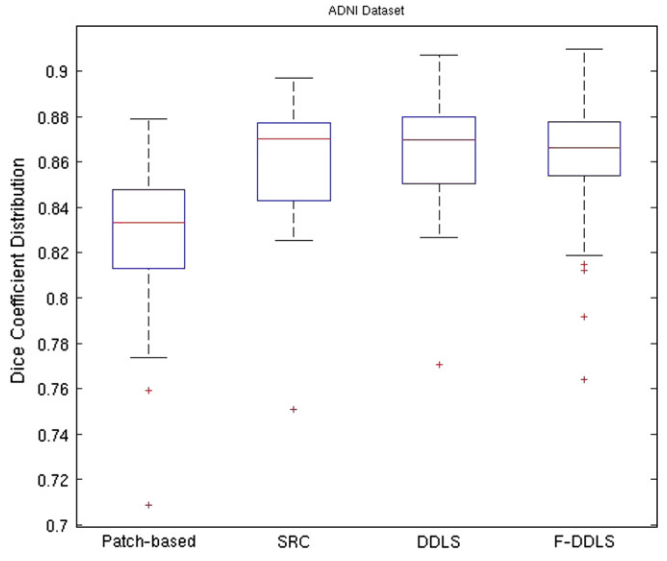

Fig. 10. Comparison of Dice Coefficient Distribution of four methods on 41 AD subjects. The results of F-DDLS were obtained by randomly selecting 30 healthy subjects for training. The other results were obtained by using the 10 most similar atlases in a leave-one-out procedure. The Patch-based and SRC methods were implemented by using a patch size of $7 \times 7 \times 7$ voxels and a search volume of $9 \times 9 \times 9$ voxels. The DDLS and F-DDLS methods were implemented by using a sampling step size of 3, a patch size of $5 \times 5 \times 5$ voxels and a search volume of $7 \times 7 \times 7$ voxels.

Computational time

We implemented the DDLS and SRC methods in MATLAB 7.13.0 using C/MEX code. The SPAMS software (http://spams-devel.gforge.inria.fr) was used for dictionary learning and sparse coding. Since the patch-based method is not open source, we used our version in a C++ implementation. The experiments were carried out using a single core of an Intel Core i7-2600 processor at 3.4 GHz with 8 GB of RAM. It took approximately 10 min for segmenting one subject by using the patch-based method with a patch size of $7 \times 7 \times 7$, a neighborhood size of $9 \times 9 \times 9$ and 10 similar atlases. The SRC method took around 40 min for segmenting one subject with the same parameter settings. For the proposed DDLS method, it took 3–6 min to segment one subject with the suggested parameter settings (a sampling step size of 3, a dictionary size of 256, a patch size of $5 \times 5 \times 5$, a neighborhood size of $7 \times 7 \times 7$ and 10 similar atlases). However, if one wants to achieve more accurate results by learning more dictionaries with a larger size, it will be more time consuming. In addition, if fixed dictionaries are trained offline, it only takes less than 1 min for segmenting one subject in the testing stage with less than a 1.5%-drop in Dice overlap compared to the results by using the DDLS method.

Discussion and conclusion

In this work we developed a novel approach for the segmentation of subcortical brain structures. Dictionary learning and sparse coding techniques have been proposed for the segmentation of brain MRI images. In contrast to other methods that rely on intensity similarities, the proposed method is based on the minimization of patch reconstruction errors. Dictionaries and classifiers are learned from atlases in one framework simultaneously. The learned dictionaries can be used for patch reconstruction and the corresponding classifiers can be used for label estimation. The proposed approach belongs to supervised learning methods by exploiting the discriminative information in the patch library extracted from atlases. To the best of our knowledge, discriminative dictionary learning has never been used in subcortical brain segmentation. The evaluation on hippocampus extraction of 202 ADNI images and 80 healthy ICBM subjects demonstrates the accuracy and robustness of the proposed method. The highest median Dice coefficient is 0.879 on ADNI dataset and 0.890 on ICBM dataset, which is competitive compared with state-of-the-art methods.

In order to reduce the computational burden of the dictionary learning process, we combined the online algorithm (Mairal et al., 2009a) with the discriminative dictionary learning approach (Zhang and Li, 2010). We also used a sampling strategy to learn the dictionaries so that the runtime of training will be shortened significantly. The dictionary and classifier related to one target voxel are learned from patches extracted in a local search volume across the atlases. Therefore, neighborhood voxels share most of template patches and will have very similar dictionaries and classifiers. This means that neighborhood voxels

Table 2
Median Dice overlaps for 80 ICBM subjects. The numbers in bold represent the highest Dice overlaps among different methods. The results of F-DDLS* were obtained by randomly selecting 40 atlases for training and using the remaining 40 subjects for evaluation. The experiment was repeated 10 times. The other results were obtained by using the most similar 20 atlases in a leave-one-out procedure. The Patch-based and SRC methods were implemented by using a patch size of $7 \times 7 \times 7$ voxels and a search volume of $9 \times 9 \times 9$ voxels. The DDLS and F-DDLS methods were implemented by using a sampling step size of 3, a patch size of $5 \times 5 \times 5$ voxels and a search volume of $7 \times 7 \times 7$ voxels. The difference between Patch-based and DDLS is statistically significant with $p < 0.001$ on Students two-tailed paired *t*-test.

Method	Right hippocampus	Left hippocampus	Whole hippocampus
Patch-based	0.882 (0.026)	0.882 (0.025)	0.882 (0.022)
SRC	0.888 (0.023)	**0.889 (0.021)**	0.888 (0.019)
DDLS	**0.892 (0.024)**	0.887 (0.020)	**0.890 (0.019)**
F-DDLS*	0.888 (0.027)	0.886 (0.025)	0.887 (0.022)

Table 3
Median Dice overlaps for 202 ADNI subjects. The numbers in bold represent the highest Dice overlaps among different methods. The results of F-DDLS* were obtained by randomly selecting 30 atlases for training and using the remaining 172 subjects for evaluation. The other results were obtained by using the 10 most similar atlases in a leave-one-out procedure. The Patch-based and SRC methods were implemented by using a patch size of $7 \times 7 \times 7$ voxels and a search volume of $9 \times 9 \times 9$ voxels. The DDLS and F-DDLS methods were implemented by using a sampling step size of 3, a patch size of $5 \times 5 \times 5$ voxels and a search volume of $7 \times 7 \times 7$ voxels. The difference between Patch-based and DDLS is statistically significant with $p < 0.001$ on Student's two-tailed paired *t*-test.

Method	Right hippocampus	Left hippocampus	Whole hippocampus
Patch-based	0.848 (0.032)	0.842 (0.029)	0.844 (0.027)
SRC	**0.873 (0.027)**	0.869 (0.026)	0.871 (0.022)
DDLS	0.872 (0.027)	**0.872 (0.031)**	**0.872 (0.024)**
F-DDLS*	0.865 (0.042)	0.859 (0.048)	0.864 (0.035)

can share dictionaries and classifiers. This may explain why the proposed sampling strategy still keeps almost the same accuracy, while generating a much faster implementation. The proposed method also uses affine registrations rather than non-rigid registrations in order to gain computational efficiency as described in the non-local means patch-based method (Coupé et al., 2011, 2012; Rousseau et al., 2011). In the end, an unseen subject can be segmented in approximately 6 min while keeping a high segmentation accuracy.

Due to different datasets for evaluation and different qualities of manual segmentations, comparison with state-of-the-art methods is always difficult. Recent works (Aljabar et al., 2009; Barnes et al., 2008; Chupin et al., 2007; Collins and Pruessner, 2010; Coupé et al., 2011; Wang et al., 2011a; Wolz et al., 2010) reported Dice values greater than 0.80 for hippocampus segmentation. Our approach can yield results comparable or more accurate than these recent published results. Our proposed method was also evaluated on the same 80 healthy ICBM subjects as those used for validations of methods proposed in Collins and Pruessner (2010), Coupé et al. (2011), and Hu et al. (2011). A median Dice value of 0.87–0.886 was achieved in these works. In comparison, our method can achieve similar or slightly improved results with a very fast implementation, especially compared to the multi-atlas method as described in Collins and Pruessner (2010).

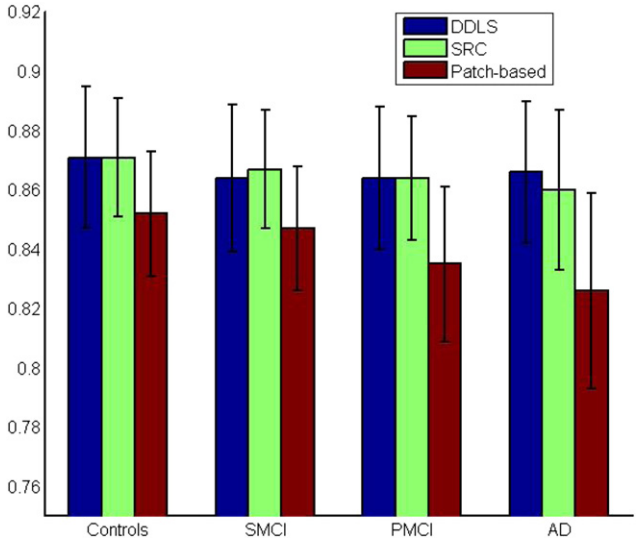

Fig. 12. Comparison of mean Dice overlaps of three methods on four different groups.

The results obtained on the ICBM dataset showed a higher Dice overlap (0.890) compared to the results on the ADNI dataset (0.879). The difference may come from the higher anatomical and scanner-based variability within the 202 ADNI subjects compared to that of the 80 healthy ICBM subjects, who were scanned on the same scanner. In addition, the improvement of the proposed DDLS method over the patch-based method is approximately 3% on the ADNI dataset compared to only 1% on the ICBM dataset in terms of Dice overlap. This difference of improvement may be caused by several factors. First, the higher anatomical variability of the ADNI dataset compared to the ICBM dataset might make segmentations more challenging. Second, the preprocessing pipelines involved for both datasets were not similar. For the ADNI dataset,

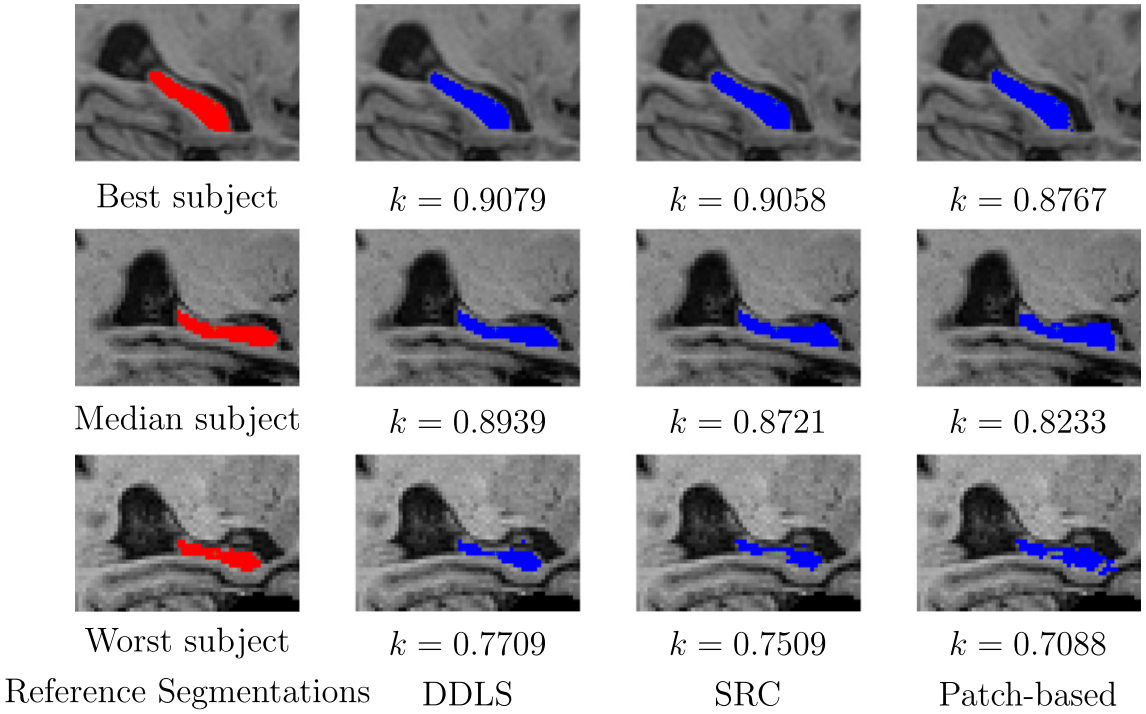

Fig. 11. Method comparison. Segmentation results were obtained by DDLS, SRC and the patch-based method for the subjects from ADNI dataset with the best, a median and the worst Dice coefficients.

the used processing may be less optimal than the pipeline used for the ICBM dataset. A less accurate intensity normalization might explain the lower performance of the nonlocal patch-based method on the ADNI dataset. However, this result highlights the robustness of the proposed methods face of intensity normalization issues compared to patch-based method. Contrary to the patch-based method, in the proposed methods the patches in the template library were normalized before used for leaning dictionaries. Finally, another reason may be that the accuracy of the proposed DDLS method on ICBM dataset may reach an upper bound because of the rater variability of manual labels (Aljabar et al., 2009), resulting in a smaller improvement on this group.

In recent work (Shu et al., 2012; Wang et al., 2011b), sparse coding techniques were also used for medical image segmentation. These methods directly used training patches as dictionaries, which are similar to the proposed SRC approach except the label fusion strategy. Weighted voting was used to fuse the labels in these approaches while the reconstruction error was used in the proposed SRC method. A further experiment was performed to compare these two different label fusion strategies. The median Dice overlaps using weighted voting as the label fusion strategy are 0.870 (0.023) on the ADNI dataset and 0.887 (0.019) on the ICBM dataset respectively, which are nearly identical to the results by using reconstruction error as the label fusion strategy (0.871 (0.022) and 0.888 (0.019) respectively). These results show that both weighted voting and reconstruction error can be chosen as reliable label fusion strategies. However, it should be noted that weighted voting cannot be used to estimate the labels in the proposed DDLS and F-DDLS approaches because the corresponding labels of the atoms in the dictionaries are unknown.

Although the proposed method can produce accurate results in a very efficient way, there are several aspects that may improve the proposed method. (1) First, the proposed approach will be improved if one discriminative dictionary with a larger size is learned for every voxel without using the sampling strategy. However, this will increase the computational cost significantly. (2) The segmentation accuracy may be improved if more complicated classifiers rather than linear classifiers are learned (Mairal et al., 2008, 2009b). Moreover, the discriminative information in the sparse coding coefficients could be also exploited and added to the objective function (Yang et al., 2011a) to improve the segmentation accuracy, although this may lead to a complicated optimization process of dictionary learning. (3) The use of non-registration instead of affine registration may also improve the segmentation results as reported in Fonov et al. (2012) and Rousseau et al. (2011). (4)Segmentation accuracy may be improved by using intensity models (Lotjonen et al., 2010; Wolz et al., 2009) or a learning-based method (Wang et al., 2011a) to correct systematic errors of the proposed method.

In Section 3.4, we also used a fixed subgroup of subjects as atlases to learn dictionaries and classifiers. The aim of this experiment was trying to learn a new format of 'atlases' and 'labels'. The dictionaries are learned from atlas images, which will contain the information of atlas images and can then be considered as new 'atlases'. Also, the corresponding classifiers contain the prior information of reference segmentations and can then be considered as new 'labels'. It may take several days to learn very good representative dictionaries and optimal discriminative classifiers offline. Once learned, it is possible to segment one target image very quickly (less than 1 min) by using the new format of 'atlases' (dictionaries) and 'labels' (classifiers). Although the median Dice value by using F-DDLS strategy drops slightly (less than 1.5%) compared to the results by using the DDLS method, it indicates that this may be a very potential direction for human brain labeling in future work.

When comparing the DDLS and F-DDLS methods, the DDLS method can achieve a better performance because it selects the most similar atlases for segmentation while in the F-DDLS method, atlases are randomly selected for segmentation, which indicates that the proposed method will have a better performance if similar types of data are used. However, the F-DDLS still achieves promising segmentation results. This means that this approach is able to segment images reliably without explicitly choosing atlases similar to the test data. The further experiment that was validated on different groups of datasets shows that the proposed F-DDLS method is able to generalize from one type of data to morphologically different datasets. The main reason may be that the proposed method learns dictionaries at a very localized level rather than a global one. However, it should be noted that the dictionaries were trained on the same imaging modality (T1 images) with the same manual segmentation protocol in our work. If the segmentation protocol, acquisition protocol or imaging modality changes, the dictionaries should be retrained.

Future research will focus on the extension of the proposed approach on the segmentation of multiple structures of brain MR images. In this paper, we just applied the proposed method to the segmentation of the hippocampus. However, it is possible to extend our proposed method to segment multiple structure by adding the corresponding label information to the label vector H in Eq. (8). In addition, we plan to apply the proposed method to the measurement of hippocampal atrophy and patient classification as proposed in Coupé et al. (2012).

Acknowledgments

This project is partially funded under the 7th Framework Programme by the European Commission (http://cordis.europa.eu/ist/) and the China Scholarship Council.

We wish to acknowledge the Montreal Neurological Institute's Brain Imaging Center and the team of Dr. Alan Evans for making MRI data available for this project. These data were collected under the auspices of the International Consortium for Brain Imaging (ICBM) project, PI: J Mazziotta (NIH-9P01EB0011955-11). We also thank Prof. Louis Collins for his helpful discussions and comments on our work.

The ADNI data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI; Principal Investigator: Michael Weiner; NIH grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and through generous contributions from the following: Pfizer Inc., Wyeth Research, Bristol-Myers Squibb, Eli Lilly and Company, GlaxoSmithKline, Merck & Co. Inc., AstraZeneca AB, Novartis Pharmaceuticals Corporation, Alzheimer's Association, Eisai Global Clinical Development, Elan Corporation plc, Forest Laboratories, and the Institute for the Study of Aging, with participation from the U.S. Food and Drug Administration. Industry partnerships are coordinated through the Foundation for the National Institutes of Health. The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory of Neuroimaging at the University of California, Los Angeles.

Conflict of interest

We declare that we have no conflict of interest.

Appendix A. The Alzheimer's Disease Neuroimaging Initiative

The Alzheimer's Disease Neuroimaging Initiative (ADNI) was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, M.D., VA

Medical Center and University of California — San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research — approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years, and 200 people with early AD to be followed for 2 years. For up-to-date information see www.adni-info.org.

Appendix B. Derivation of discriminative dictionary learning

For solving Eq. (8), we rewrite the equation as:

$$\langle D, W, \alpha \rangle = \underset{\substack{D,W,\alpha \\ \text{subject to } \|\alpha\|_0 \le T}}{argmin} \left\| \begin{pmatrix} P_L \\ \sqrt{\beta_1} H \end{pmatrix} - \begin{pmatrix} D \\ \sqrt{\beta_1} W \end{pmatrix} \alpha \right\|_2^2. \tag{B.1}$$

Let $\tilde{D} = (D^t, \sqrt{\beta_1} W^t)^t$, $\tilde{P}_L = (P_L{}^t, \sqrt{\beta_1} H^t)^t$. Each column of the input signal $\tilde{P}_L$ will thus include the original patch and its corresponding label information. Each atom of dictionary $\tilde{D}$ is always normalized. We also use the l_1 norm to achieve a sparse solution for α in the dictionary optimization process. Eq. (B.1) can be rewritten as:

$$\left\langle \tilde{D}, \alpha \right\rangle = \underset{\tilde{D},\alpha}{argmin} \left\| \tilde{P}_L - \tilde{D}\alpha \right\|_2^2 + \beta_2 \|\alpha\|_1 \tag{B.2}$$

and can be efficiently solved by the online dictionary learning technique described in Mairal et al. (2009a). This online approach draws one patch from the patch library at a time for updating the dictionary. In our implementation, a mini-batch strategy (Mairal et al., 2009a), which draws a few patches at each iteration rather than a single patch, was used to improve the convergence speed of online learning.

After online dictionary learning, we obtain a dictionary and a classifier for every target voxel. The corresponding dictionary and classifier of the target patch p_t can be represented as:

$$\tilde{D}_t = \left\{ \begin{pmatrix} \tilde{d}_1 \\ \sqrt{\beta_1} \tilde{w}_1 \end{pmatrix}, \begin{pmatrix} \tilde{d}_2 \\ \sqrt{\beta_1} \tilde{w}_2 \end{pmatrix}, \ldots, \begin{pmatrix} \tilde{d}_K \\ \sqrt{\beta_1} \tilde{w}_K \end{pmatrix} \right\} \tag{B.3}$$

where the learned dictionary and the corresponding classifier parameters are normalized jointly. As a result, we cannot use these dictionaries and the classifiers directly for labeling. However, the desired dictionary $\hat{D}_t$ and the classifier parameters $\hat{W}_t$ can be computed from $\tilde{D}_t$:

$$\begin{aligned} \hat{D}_t &= \left\{ \frac{\tilde{d}_1}{\|\tilde{d}_1\|_2}, \frac{\tilde{d}_2}{\|\tilde{d}_2\|_2}, \ldots, \frac{\tilde{d}_K}{\|\tilde{d}_K\|_2} \right\} \\ \hat{W}_t &= \left\{ \frac{\tilde{w}_1}{\|\tilde{d}_1\|_2}, \frac{\tilde{w}_2}{\|\tilde{d}_2\|_2}, \ldots, \frac{\tilde{w}_K}{\|\tilde{d}_K\|_2} \right\} \end{aligned}. \tag{B.4}$$

The proofs of Eq. (B.4) are available in Zhang and Li (2010).

References

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

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

Babalola, K., Patenaude, B., Aljabar, P., Schnabel, J., Kennedy, D., Crum, W., Smith, S., Cootes, T., Jenkinson, M., Rueckert, D., 2009. An evaluation of four automatic methods of segmenting the subcortical structures in the brain. Neuroimage 47 (4), 1435–1447.

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

Chupin, M., Mukuna-Bantumbakulu, A., Hasboun, D., Bardinet, E., Baillet, S., Kinkingnéhun, S., Lemieux, L., Dubois, B., Garnero, L., 2007. Anatomically constrained region deformation for the automated segmentation of the hippocampus and the amygdala: method and validation on controls and patients with Alzheimer's disease. Neuroimage 34 (3), 996–1019.

Collins, D., Holmes, C., Peters, T., Evans, A., 1995. Automatic 3-D model-based neuroanatomical segmentation. Hum. Brain Mapp. 3 (3), 190–208.

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

Coupé, P., Eskildsen, S., Manjón, J., Fonov, V., Collins, D., 2012. Simultaneous segmentation and grading of anatomical structures for patient's classification: application to Alzheimer's disease. Neuroimage 59 (4), 3736–3747.

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

Eskildsen, S., Coupé, P., Fonov, V., Manjón, J., Leung, K., Guizard, N., Wassef, S., Østergaard, L., Collins, D., 2011. BEaST: brain extraction based on nonlocal segmentation technique. Neuroimage 59 (3), 2362–2373.

Fonov, V., Coupé, P., Styner, M., Collins, L., et al., 2012. Automatic lateral ventricle segmentation in infant population with high risk of autism. 18th Annual Meeting of the Organization for Human Brain Mapping (OHBM).

Heckemann, R., Hajnal, J., Aljabar, P., Rueckert, D., Hammers, A., 2006. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. Neuroimage 33 (1), 115–126.

Hsu, Y., Schuff, N., Du, A., Mark, K., Zhu, X., Hardin, D., Weiner, M., 2002. Comparison of automated and manual MRI volumetry of hippocampus in normal aging and dementia. J. Magn. Reson. Imaging 16 (3), 305–310.

Hu, S., Coupé, P., Pruessner, J., Collins, D., 2011. Appearance-based modeling for segmentation of hippocampus and amygdala using multi-contrast MR imaging. Neuroimage 58 (2), 549–559.

Jack, C., Bernstein, M., et al., 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Reson. Imaging 27 (4), 685–691.

Jiang, Z., Lin, Z., Davis, L., 2011. Learning a discriminative dictionary for sparse coding via label consistent K-SVD. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1697–1704.

Khan, A., Cherbuin, N., Wen, W., Anstey, K., Sachdev, P., Beg, M., 2011. Optimal weights for local multi-atlas fusion using supervised learning and dynamic information (SuperDyn): validation on hippocampus segmentation. Neuroimage 56 (1), 126–139.

Kittler, J., 1998. Combining classifiers: a theoretical framework. Pattern Anal. Appl. 1 (1), 18–27.

Lotjonen, J., Wolz, R., Koikkalainen, J., Thurfjell, L., Waldemar, G., Soininen, H., Rueckert, D., et al., 2010. Fast and robust multi-atlas segmentation of brain magnetic resonance images. Neuroimage 49 (3), 2352–2365.

Mairal, J., Bach, F., Ponce, J., 2012. Task-driven dictionary learning. IEEE Trans. Pattern Anal. Mach. Intell. 34 (4), 791–804.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., 2009a. Online dictionary learning for sparse coding. Proceedings of the 26th Annual International Conference on Machine Learning. ACM, pp. 689–696.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A., 2008. Discriminative learned dictionaries for local image analysis. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A., 2009b. Supervised dictionary learning. Adv. Neural Inf. Process. Syst. 21, 1033–1040.

Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., 1995. A probabilistic atlas of the human brain: theory and rationale for its development: the international consortium for brain mapping (ICBM). Neuroimage 2 (2), 89–101.

Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L., 2005. The Alzheimer's Disease Neuroimaging Initiative. Neuroimaging Clin. N. Am. 15 (4), 869–877.

Nyul, L., Udupa, J., 2000. Standardizing the MR image intensity scales: making MR intensities have tissue-specific meaning. Medical Imageing: Image Display and Visualization in Proceedings of SPIE 1: , 21, pp. 496–504.

Pruessner, J., Li, L., Serles, W., Pruessner, M., Collins, D., Kabani, N., Lupien, S., Evans, A., 2000. Volumetry of hippocampus and amygdala with high-resolution MRI and three-dimensional analysis software: minimizing the discrepancies between laboratories. Cereb. Cortex 10 (4), 433–442.

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

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

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

Shu, L., Gao, Y., Shen, D., 2012. Sparse patch based prostate segmentation in CT images. Med. Image Comput. Comput.-Assist. Interv. 2012, 385–392.

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

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

Van der Lijn, F., den Heijer, T., Breteler, M., Niessen, W., 2008. Hippocampus segmentation in MR images using atlas registration, voxel classification, and graph cuts. Neuroimage 43 (4), 708–720.

Wang, H., Das, S., Suh, J., 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 (3), 968–985.

Wang, H., Suh, J., Das, S., Pluta, J., Altinay, M., Yushkevich, P., 2011b. Regression-based label fusion for multi-atlas segmentation. 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1113–1120.

Wang, H., Suh, J., Pluta, J., Altinay, M., Yushkevich, P., 2011c. Optimal weights for multi-atlas label fusion. Information Processing in Medical Imaging. Springer, pp. 73–84.

Wang, H., Yushkevich, P., 2012. Dependency prior for multi-atlas label fusion. IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI), pp. 892–895.

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

Wolz, R., Aljabar, P., Hajnal, J., Hammers, A., Rueckert, D., et al., 2010. LEAP: learning embeddings for atlas propagation. Neuroimage 49 (2), 1316–1325.

Wolz, R., Aljabar, P., Rueckert, D., Heckemann, R., Hammers, A., 2009. Segmentation of subcortical structures and the hippocampus in brain MRI using graph-cuts and subject-specific a-priori information. IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI), pp. 470–473.

Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y., 2009. Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. 31 (2), 210–227.

Yang, A., Sastry, S., Ganesh, A., Ma, Y., 2010. Fast l1-minimization algorithms and an application in robust face recognition: a review. International Conference on Image Processing (ICIP), pp. 1849–1852.

Yang, M., Zhang, L., Feng, X., Zhang, D., 2011a. Fisher discrimination dictionary learning for sparse representation. IEEE International Conference on Computer Vision (ICCV), pp. 543–550.

Yang, M., Zhang, L., Yang, J., Zhang, D., 2011b. Robust sparse coding for face recognition. IEEE International Conference on Computer Vision (ICCV), pp. 625–632.

Zhang, Q., Li, B., 2010. Discriminative K-SVD for dictionary learning in face recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2691–2698.

Zhang, S., Zhan, Y., Metaxas, D.N., 2012. Deformable segmentation via sparse representation and dictionary learning. Med. Image Anal. 1385–1396.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Methodol. 67 (2), 301–320.