Medical Image Analysis 18 (2014) 808–818

Contents lists available at ScienceDirect

Medical Image Analysis

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

Multiple instance learning for classification of dementia in brain MRI

Tong Tong [a,*], Robin Wolz [a], Qinquan Gao [a], Ricardo Guerrero [a], Joseph V. Hajnal [b], 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] 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:
Received 29 July 2013
Received in revised form 9 April 2014
Accepted 16 April 2014
Available online 5 May 2014

Keywords:
Structural MR imaging
Classification
Multiple instance learning
Alzheimer's disease

ABSTRACT

Machine learning techniques have been widely used to detect morphological abnormalities from structural brain magnetic resonance imaging data and to support the diagnosis of neurological diseases such as dementia. In this paper, we propose to use a multiple instance learning (MIL) method in an application for the detection of Alzheimer's disease (AD) and its prodromal stage mild cognitive impairment (MCI). In our work, local intensity patches are extracted as features. However, not all the patches extracted from patients with dementia are equally affected by the disease and some of them may not be characteristic of morphology associated with the disease. Therefore, there is some ambiguity in assigning disease labels to these patches. The problem of the ambiguous training labels can be addressed by weakly supervised learning techniques such as MIL. A graph is built for each image to exploit the relationships among the patches and then to solve the MIL problem. The constructed graphs contain information about the appearances of patches and the relationships among them, which can reflect the inherent structures of images and aids the classification. Using the baseline MR images of 834 subjects from the ADNI study, the proposed method can achieve a classification accuracy of 89% between AD patients and healthy controls, and 70% between patients defined as stable MCI and progressive MCI in a leave-one-out cross validation. Compared with two state-of-the-art methods using the same dataset, the proposed method can achieve similar or improved results, providing an alternative framework for the detection and prediction of neurodegenerative diseases.

1. Introduction

The aetiology of Alzheimer's disease (AD) is the most commonly responsible for clinical dementia worldwide. Its progression leads to a gradual decline of memory and cognitive functions. The prevalence of AD is predicted to quadruple in the next four decades (Brookmeyer et al., 2007). However, no drug or treatment has so far been reported to be able to stop the progress of AD and it remains difficult to predict whether individuals will develop AD. There is a critical need to develop biomarkers for the early diagnosis of AD and measuring the outcomes of clinical drug trials (Clark et al., 2007). Although there is currently no cure for AD, there are some medications that can delay the onset of some symptoms such as memory loss, confusion, and cognitive problems (Yiannopoulou and Papageorgiou, 2013). Diagnosing AD early would allow doctors to treat patients sooner, which can then limit the devastating physical, psychological impact on patients and their relatives and reduce the economic burden on society. Mild cognitive impairment (MCI) is an intermediate stage between normal cognition and clinical dementia. Individuals with MCI have been reported to progress to clinical dementia at a rate of 10–15% annually (Grundman et al., 2004). Research on identifying MCI individuals who will progress to clinical dementia has received increasing attention in recent years (Wolz et al., 2011; Coupé et al., 2012; Wee et al., 2012b; Liu et al., 2013; Gray et al., 2013).

Different imaging techniques, such as structural magnetic resonance imaging (MRI) (Wolz et al., 2011; Coupé et al., 2012; Liu et al., 2013), functional MRI (Pihlajamäki and Sperling, 2008; Wee et al., 2011), fluorodeoxyglucose positron emission tomography (FDG-PET) (Herholz et al., 2002; Gray et al., 2012) and diffusion tensor imaging (DTI) (Wee et al., 2012b; Keihaninejad et al., 2013), have been used to derive image-based biomarkers for AD.

* Corresponding author. Tel.: +44 20 7852 1982.
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 (http://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.

http://dx.doi.org/10.1016/j.media.2014.04.006
1361-8415/

Studies have shown that the combination of biomarkers from different imaging modalities (MRI, FDG-PET, DTI, fMRI) can provide complementary information of AD pathology and thus improve the classification accuracy (Zhang et al., 2011; Hinrichs et al., 2011; Wee et al., 2012b; Gray et al., 2013). In comparison to DTI, fMRI or FDG-PET, structural MRI is the most standardized and the most widely available imaging modality in clinical practice. In addition, MRI examinations can provide an opportunity to track different clinical phases of AD (Jack et al., 2013). Therefore, we evaluated our method using structural MR images. However, multiple datasets could also be acquired from different imaging modalities for developing different biomarkers of AD.

Several types of features can be derived from structural MRI for classification, such as gray matter density maps (Cuingnet et al., 2011; Liu et al., 2012a), cortical thickness (Cho et al., 2012; Wee et al., 2012a; Eskildsen et al., 2013) as well as volume and shape measures (Gerardin et al., 2009; Wolz et al., 2010). The number of training images is typically small in comparison with the high dimensionality of the voxel-wise features. Therefore, a feature selection step is necessary to tackle the problem of overfitting. Feature selection has been shown to improve the classification accuracy, but it depends on the adopted approaches (Chu et al., 2011). To reduce the feature space and select the discriminative features, statistical approaches (Yoon et al., 2007; Chu et al., 2011; Wee et al., 2012a) or sparse regression methods (Ghosh and Chinnaiyan, 2005; Liu et al., 2012b) are often used. Another popular method is to segment the whole brain into multiple anatomical (Gray et al., 2012) or discriminative (Fan et al., 2007) regions and then extract regional features such as volume or shape measures for classification. It should be noted that the features extracted from neuroimaging data are not isolated and exhibit high correlations (Chu et al., 2011). Considering the relationships among these features, tree-guided sparse coding methods (Liu et al., 2012b) or re-sampling schemes using Elastic Net (Janousova et al., 2012) has been recently proposed. These approaches can select voxel-wise features in meaningful brain regions, which may be related to pathology.

The features derived from MRI can be extracted from very local regions or the whole brain. At the voxel level, intensities or gray matter densities can be directly used in classification (Cuingnet et al., 2011; Vounou et al., 2012). At the whole image level, similarities between images can be used to derive features (Wolz et al., 2012). However, the structural changes induced in the early stages of AD have been observed to occur in small local regions rather than isolated voxels or the whole brain (Hinrichs et al., 2009). Patches represent features at an intermediate scale between the voxel level and the image level, which can capture disease-induced changes in local regions. Recent approaches (Coupé et al., 2012; Liu et al., 2013) utilize local intensity patterns within patches to capture the local structural information for AD classification. In these approaches, patches from patients with AD are used as positive samples and patches from healthy subjects are regarded as negative samples for training. However, patches are relatively small regions in brain images and not all patches in the brain are characteristic of changes associated with pathology. For example, patches in close vicinity of the hippocampus are more likely to be affected by AD while patches in homogeneous regions may not be affected. This is illustrated in Fig. 1. In addition, different types of dementias have different aetiologies. This means that some patches may be affected by other aetiologies such as cerebrovascular disease rather than AD. Therefore, not all patches from patients necessarily represent positive training samples. This means that there is some ambiguity in assigning disease labels to the training patches extracted from patients. One solution to this problem is to use a weakly supervised method such as multiple instance learning (MIL) (Maron and Lozano-Pérez, 1998), which can learn classifiers from ambiguously labeled training data. Although MIL have been successfully applied to different applications in computer vision (Babenko et al., 2009) and recently in medical imaging (Bi and Liang, 2007; Xu et al., 2012), to the best of our knowledge, it has not been used in the context of classification of neurological diseases. In this paper, we propose to use MIL for the classification of AD and to address the problem of ambiguous patch labels. Specifically, each image is regarded as a bag; the patches extracted from the images are thus treated as inter-correlated instances in the bags. MIL is then used to learn a bag-level classifier to predict the bag labels of unseen images and therefore classify the subjects.

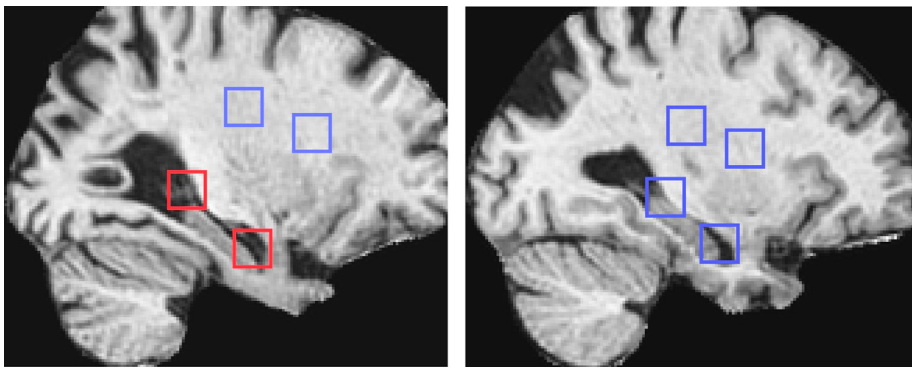
Most existing approaches utilize the intensity values of patches for classification. The relationships among patches are usually ignored since the patches are treated as independently and identically distributed. However, patches from the same subject are rarely independent and often exhibit shared information. This information across patches can convey information about the inherent structure of the images, which may be helpful for disease classification. In recent works, correlated features are extracted to exploit the relationships among patches (Liu et al., 2013) or ROIs (Wee et al., 2012a) of the same subject, which has been shown to improve the classification accuracy. In our work, a graph is constructed from each image in order to investigate the relationships among patches and to exploit the inherent structural information of each image. After that, a graph kernel, which utilizes both the intensity values and the relationships of the extracted patches, is used to distinguish the positive and negative bags. Finally, a bag-level classifier is trained via a kernel machine.

A preliminary version of the presented framework has been published as a conference paper (Tong et al., 2013). The major difference in this work is that we adopted a more robust feature selection method as proposed in Janousova et al. (2012). In addition, an extended evaluation on the whole brain is presented and more detailed comparisons with state-of-the-art methods are also provided. The remainder of this paper is organized as follows: The demographic information of the image dataset in preparation of this article is introduced in Section 2.1. This is followed by a description of the preprocessing pipeline of these images in Section 2.3 and a description on how patches are extracted from the images to form corresponding bags in Section 2.4. We will then introduce the methodology of MIL and how we apply it to the classification of AD in Section 2.5. Performance of the proposed method has been evaluated using 834 subjects from the ADNI study. In Section 3, the influence of different parameters are studied and the performance of the proposed method is also compared with state-of-the-art techniques. The strengths and weaknesses of the proposed method are analyzed in the discussion section and finally we conclude the paper in Section 5.

2. Materials and methods

2.1. Subjects

Data used in the preparation of this article were obtained from the ADNI database (adni.loni.ucla.edu). The 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, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early 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.



(a) patient with AD (b) control subject

Fig. 1. Example of different bags. (a) Positive bag (AD patient); (b) negative bag (control subject); The red boxes and blue boxes represent positive patches and negative patches respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In our work, 834 baseline MR scans at 1.5 T were downloaded in July 2011 from the ADNI database for evaluation, which consist of 231 cognitively normal (CN), 238 stable MCI (SMCI), 167 progressive MCI (PMCI) and 198 AD. Subjects in the MCI group were classified as PMCI if the subjects converted to AD during a 3 years follow-up. Those who did not convert after a 3 years follow-up period were classified as SMCI. These four groups are the same as the subjects used in Wolz et al. (2011) and in Coupé et al. (2012). The demographics of these 834 subjects are shown in Table 1.

2.2. Overview of methods

An overview of the proposed method mi-Graph (multiple instance-Graph) is shown in Fig. 2. In the proposed mi-Graph, patches are extracted from preprocessed images and regarded as features. The labels of the training images are known while the labels of the training patches extracted from patients are unknown. Images are regarded as bags and patches are treated as instances in the bags. As shown in Fig. 1, if the bag contains at least one positive patch related to disease changes, the bag is labeled as positive; otherwise, the bag is labeled as negative (in this case all the patches in the bag are negative). Therefore, images from patients and controls can be regarded as positive and negative bags respectively. Then, this multiple instance learning problem can be solved using a number of different approaches (Fu et al., 2011). The final goal is to learn a bag-level classifier to label unseen bags (i.e. images).

In this paper, we propose to use a graph-based multiple instance learning method (Zhou et al., 2009) for learning the bag-level classifier. In this method, a graph is constructed for each image. The patches in each bag are treated as nodes in the corresponding graph and edges between different nodes are established according to the relationships between patches. The graphs can represent the appearances of patches and reflect the relationships among the patches extracted from the same subject. Since some patches extracted from patients may be affected by AD while the patches extracted from controls are not affected by AD, the resulting graphs are expected to be different between different groups. A graph kernel is defined for distinguishing the positive and negative bags. Finally, a bag-level classifier can be learned using a kernel machine such as support vector machine (SVM). Accordingly, there are four major steps in the proposed framework: Image preprocessing, extraction of patches, computation of graph kernels and classifier training. In the following, we will present the details of these steps.

Table 1
Demographic information describing dataset used in this study. Disease statuses of MCI subjects are defined after a 3 years follow up.

Group	Number	Age	Male (%)	MMSE[a]	CDR[b]
CN	231	76.0 ± 5.0	52	29.1 ± 1.0	0 ± 0
SMCI	238	74.9 ± 7.8	66	27.3 ± 1.8	0.49 ± 0.05
PMCI	167	74.6 ± 7.0	62	26.6 ± 1.7	0.50 ± 0
AD	198	75.7 ± 7.7	52	23.3 ± 2.0	0.75 ± 0.25

[a] MMSE: Mini-Mental State Examination (Score 0–30).
[b] CDR: Clinical Dementia Rating (0.5 – very mind; 1 – mild; 2 – moderate; 3 – severe).

2.3. Image processing

The T1-weighted MR brain images were preprocessed by the standard ADNI pipeline as described in Jack and Bernstein (2008), which includes post-acquisition correction of gradient warping, B1 non-uniformity correction, intensity non-uniformity correction and phantom-based scaling correction. All the images were skull-stripped using the method proposed in Leung et al. (2011). After that, non-rigid registration was performed to align all images to the MNI152 template space using non-rigid registration based on B-spline free-form deformation (Rueckert et al., 1999) with a final control point spacing of 2.5 mm. The approach proposed in Nyúl and Udupa (1999) was used to normalize the intensities between the subjects and the template. After preprocessing, all the images are spatially normalized and the intensities are homogeneous across the images.

2.4. Extraction of patches

In the proposed method, each image is regarded as a bag and patches are exacted from each image to form the bag. For each image, the set of all possible patches is equal to its dimensionality. This means that the total number of patches M is extremely high and only $K \ll M$ patches are extracted to form the corresponding bag. A simple way to extract K patches is to randomly select patches from the image. However, this cannot guarantee that there are positive patches in positive bags and is not optimal for the extraction of patches. Ideally, the extracted K patches should be discriminative between groups. At the same time, the patches need to be representative and should reflect information about the inherent structure of the images. In order to extract discriminative patches, we assign probabilities to different patches. The assigned probabilities should represent the discriminative ability of the corresponding patches. If patches at a specific location are highly discriminative between different groups, high probabilities will

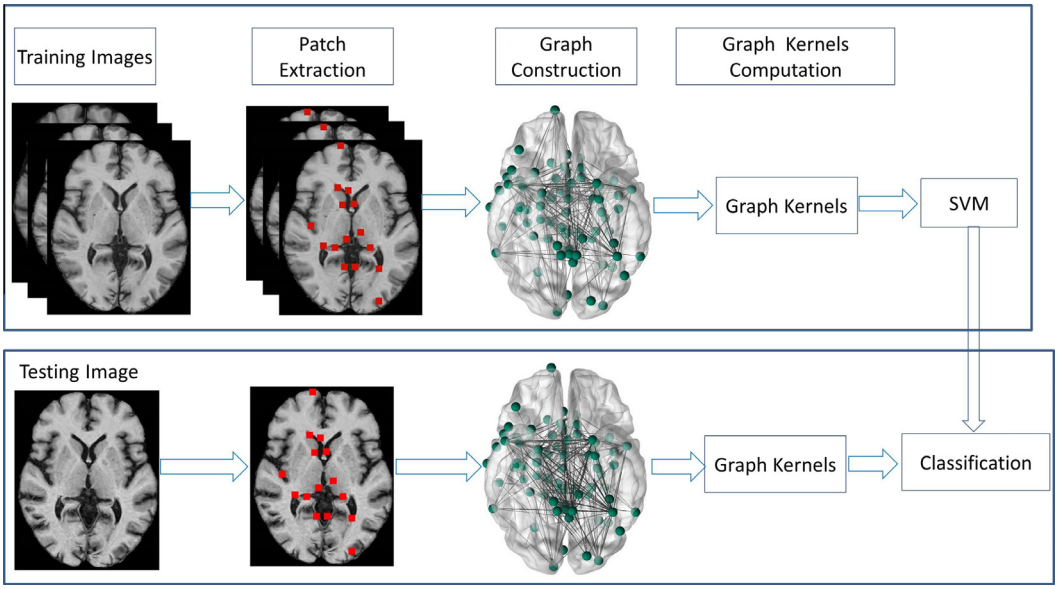


Fig. 2. Flow chart of the proposed mi-Graph method.

be assigned to these patches. The patches with high discriminative probabilities are then extracted for classification.

However, this does not mean that the more discriminative patches in the bags, the better performance the classifier achieves. The classification accuracy is also affected by the relationships among the selected patches (Zhou et al., 2009; Liu et al., 2013) since these patches are used as features. If the selected patches have large overlap, these patches will contain a large amount of redundant feature information and provide limited information about the inherent structure of the images. In our case, the discriminative patches with high probabilities are more likely to lie in contiguous regions. This will result in large overlap among patches if the patches are selected just according to the discriminative probability map. To avoid this, we defined a spatial distance threshold to control the overlap between patches. The spatial distance threshold is calculated between the center locations of the patches. Finally, K patches are extracted from each image according to the discriminative probability map and the defined distance threshold.

Different methods such as t-tests (Chu et al., 2011) or elastic net (Janousova et al., 2012) can be used to calculate the probabilities for selecting patches. In our previous work (Tong et al., 2013), t-tests were used to select K patches to form the training bags. In this paper, the elastic net method proposed in Janousova et al. (2012) was used instead of t-tests since this method can identify highly discriminative regions. In order to generate a probability map for selecting patches, a resampling scheme proposed in Janousova et al. (2012) was used. The resampling scheme repeatedly fits a sparse regression model on randomly selected subsets of the data set and keep track of voxels that are consistently selected. Finally, the average selection frequencies of voxels are treated as their selection probabilities. The probability of a voxel is defined as:

$$P_{V_j} = \frac{1}{B}\sum_{b=1}^{B} S_{V_j} \quad (1)$$

In the above equation, S_{V_j} is an indicator variable, which is set to 1 if voxel V_j is selected after one regression process. Fig. 3 shows an example of the probability map generated by using the approach proposed in Janousova et al. (2012). The selection of K patches is then performed on the probability map. The first patch is extracted at the location with the highest probability. Then, the selection probabilities of its neighboring patches within a predefined spatial distance are set to zero. As a result, the neighboring locations of the previously selected patches will not be extracted. The next patch is extracted at the location with the second highest probability. This selection step repeats until K patches are extracted. Fig. 3 also shows an example of the selected patches for the classification of CN vs AD.

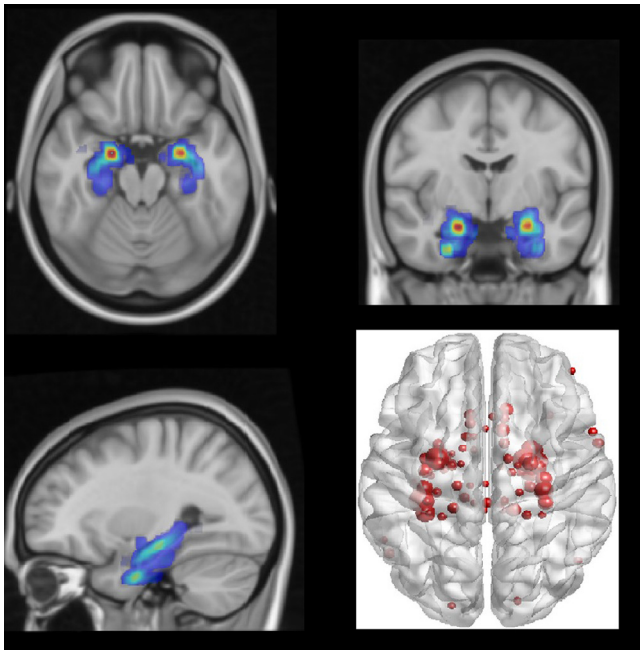


Fig. 3. Orthogonal view of the probability map for the selection of discriminative patches. The figure on the bottom right corner shows the selected 80 patches over the whole brain, which are among the most important patches for the classification of AD vs CN. The size of the red nodes indicates the discriminative probabilities of the selected patches.

2.5. Computation of graph kernels

MIL was proposed to handle the ambiguity of instance labels in positive bags (Maron and Lozano-Pérez, 1998) and has been

successfully applied to various tasks in computer vision (Andrews et al., 2002; Babenko et al., 2009). In recent years, encouraging results have also been reported in medical imaging using MIL (Bi and Liang, 2007; Lu et al., 2011; Xu et al., 2012). A method based on MIL was proposed to detect pulmonary embolisms (Bi and Liang, 2007). MIL was also adopted to detect polyps in a CAD system (Lu et al., 2011). An integrated framework was proposed to perform classification and segmentation of colon histopathology images using a context-constrained MIL (Xu et al., 2012). However, these approaches typically treat patches in the bags as independent instances and neglect their relationships. The relationships among patches, however, can provide complementary information and may be beneficial for learning strong classifiers. In our work, a graph is constructed for every image to integrate the information about the appearances of patches and the information about the relationships among patches.

Let N indicate the number of training images. After the patches are extracted from each training image, we have N training bags with K patches. The intensity values within a patch are rearranged into a feature vector and denoted as p in our work. Given a training data set $\{(B_1,y_1),\cdots,(B_i,y_i),\cdots,(B_N,y_N)\}$, where $B_i=\{p_{i1},\cdots p_{ij},\cdots p_{iK}\}$ represents K patches extracted from the image i and $y_i\in Y=\{1,0\}$ is the corresponding label of the bag B_i, the goal is to train a bag-level classifier to predict the label of the test images. The construction of the graph for each bag is quite straightforward: Similar to approaches in manifold learning (Pless and Souvenir, 2009), distance matrices are derived to construct graphs, which can capture the underlying manifold structure of the data. Each patch in the bag B_i can be viewed as a node in the graph G_i. The distance between every pair of nodes is calculated and used to define the distance matrix W^i. Various measures can be adopted to calculate the distance between nodes. Here, we simply used the squared Euclidean distance to establish the graph, but it should be noted that our method is not limited to a specific distance measure. Distances between patches are calculated using the intensity values within patches and the distance matrix W^i is defined as:

$$W^i_{au}=\|p_{ia}-p_{iu}\|_2^2 \tag{2}$$

where p_{ia} and p_{iu} are two patches in the bag B_i. The distance matrix W^i can then represent a graph that models the relationships among the patches in the bag B_i. In the resulting graph, the weight of each edge corresponds to the dissimilarity between the corresponding pair of patches within the bag.

After mapping the bags to graphs, different options can be chosen to train a classifier. For example, a k-nearest neighbor classifier can be trained by employing graph edit distance as described in Neuhaus and Bunke (2007). In this paper, we chose to define a graph kernel as proposed in Zhou et al. (2009) to capture the similarity among graphs and then train a classifier using kernel machines. Given two bags B_i and B_j which are represented as graphs with matrices W^i and W^j respectively, the graph kernel K_G is defined as

$$K_G(B_i,B_j)=\frac{\sum_{a=1}^{K}\sum_{b=1}^{K}d_{ia}d_{jb}k(p_{ia},p_{jb})}{\sum_{a=1}^{K}d_{ia}\sum_{b=1}^{K}d_{jb}} \tag{3}$$

where $d_{ia}=1/\sum_{u=1}^{K}W^i_{au}$, $d_{jb}=1/\sum_{v=1}^{K}W^j_{bv}$. In the above equation, the kernel function k is defined as

$$k(p_{ia},p_{jb})=\exp\left(-\gamma\|p_{ia}-p_{jb}\|\right) \tag{4}$$

where p_{ia} and p_{jb} are patches in bags B_i and B_j respectively. As shown in Eqs. (3) and (4), the nodes p which represent the intensities of patches and the edges w which reflect the relationships among patches are important for calculating the graph kernel K_G. Finally, the graph kernel K_G is normalized

$$\overline{K_G}(B_i,B_j)=\frac{K_G(B_i,B_j)}{\sqrt{K_G(B_i,B_i)}\sqrt{K_G(B_j,B_j)}} \tag{5}$$

2.6. Classification Using SVM

After the graph kernels are calculated, a classifier can be trained via a kernel machine. Various kernel machines such as kernel Fisher linear discriminant analysis (LDA) (Mika et al., 1999), kernel principal components analysis (PCA) (Schölkopf et al., 1997), and support vector machine (SVM) (Amari and Wu, 1999) can be used to solve the classification problem. Among them, SVM is one of the most widely used kernel machines because of its accurate classification performance (Sanchez and David, 2003). In this paper, we chose SVM to train a classifier using the computed graph kernels. In the test stage, the labels of unseen images are estimated using the learned classifier.

3. Experiments and results

The performance of the proposed mi-Graph was evaluated on different classification tasks, including CN vs AD, CN vs PMCI and SMCI vs PMCI. Experiments were performed using leave-one-out cross validation since this validation is known to be an almost unbiased estimator (Cawley and Talbot, 2004). For a fair comparison with the study in Wolz et al. (2011), we also utilized a leave 5% out cross validation as adopted in their work. There are five important parameters in our proposed method: the size of the patch, the number of the selected patches, the spatial distance threshold, the gamma in Eq. (4) and the cost parameter C in kernel SVM. The effect of parameters including the number of selected patches and the spatial patch distance threshold is analyzed in the following section. In all cases a patch size of $7\times7\times7$ voxels was used to capture local structural information as this is suggested to be a good choice in related work (Coupé et al., 2012; Liu et al., 2012a). The parameter γ in Eq. (4) was determined over the ranges $\gamma=10,5,2,1,0.5,0.1$ and a suitable one (a value of 2 for CN vs AD and 1 for SMCI vs PMCI) was chosen for all the other experiments. This was carried out because it will be computational impossible to optimize the parameter gamma using a nested leave-one-out cross validation. For training a bag-level classifier, the LIBSVM toolbox (Chang and Lin, 2011) was used and a grid-search of the cost parameter C in LIBSVM was performed over the ranges $C=10^{-5},10^{-4.5},\ldots,10^{2.5},10^{3}$ as suggested in Cuingnet et al. (2011). Finally, the performance of the proposed mi-Graph was compared with that of standard linear SVM and those of two state-of-the-art methods (Wolz et al., 2011; Coupé et al., 2012).

All the above parameters were tuned on the same groups except for the locations of selected patches, which were determined on separate groups. For classification of CN vs AD, the selection of patches was determined on the dataset of MCI groups to avoid the selection bias as described in (Kriegeskorte et al., 2009), which may lead to overestimated classification accuracy (Eskildsen et al., 2013). Also, for classification of SMCI vs PMCI, the location of patches were determined on the CN and AD groups. This was done also because of the highly computational cost when generating the probability map for selecting patches. In order to generate the probability map using the approach proposed in Janousova et al. (2012), 1000 runs of this process was repeated and the average selection frequencies are treated as probability map for selecting patches. In addition, it was found in Chu et al. (2011) that the use of a region of interest (ROI) can yield higher accuracies than using the whole brain for the classification of AD. The most frequently top ranked ROI is hippocampus. Therefore, a ROI around the hippocampus is used to restrict the selection of useful patches

in our work. This ROI (see Fig. 4) is defined around the hippocampus of the MNI152-brain T1 atlas (Proverb, 1995) followed by a dilation of 5 mm. We also evaluated the proposed mi-Graph over the whole brain and compared the results with those using the defined ROI.

3.1. Effect of parameters

Experiments were conducted to investigate the effect of the number of selected patches and the spatial distance threshold on the classification accuracy. Fig. 5 shows the classification accuracies for CN vs AD and SMCI vs PMCI when varying the number of patches and the patch distance thresholds. As can be seen from the figure, the spatial distance threshold has more effect on the classification accuracy than the number of patches. When varying the number of patches, the number of nodes in the graphs will change accordingly but the classification accuracy only slightly changes. When different spatial distance thresholds are used, the selected patches are different and the edges between nodes in the graphs change accordingly. The edges reflect the relationships among these selected patches. When the spatial distance threshold varies, the classification accuracy changes significantly, which indicates that the relationships among patches are very important for the classification task.

As shown in Fig. 5, when a patch distance threshold of 1 is used, the classification accuracy is much lower than those using other patch distance thresholds. When a patch distance threshold of 1 is adopted, the extracted patches are selected just according to their discriminative probabilities. In this case, the selected K patches have the highest discriminative probabilities of all M patches. This means that these patches are more discriminative than the K patches extracted using other patch distance thresholds. However, these K patches have significant overlap and provide limited information about the inherent structure of the images, thus leading to a low classification accuracy. When using a larger patch distance threshold (e.g. 5 or 7 voxles), the proposed mi-Graph achieves classification accuracies of 0.890 for CN vs AD and 0.704 for SMCI vs PMCI using a leave-one-out cross validation.

3.2. Whole brain vs hippocampus

We also evaluated the proposed mi-Graph using the whole brain. Patches were extracted over the whole brain to establish graphs. The results using the whole brain were compared with the results using the defined ROI around hippocampus. The optimal spatial distance thresholds were used in both classification scenarios for a fair comparison. Fig. 6 shows the results when different number of patches are selected. As can be seen from the figure, the classification accuracies using the whole brain are lower than those obtained using the defined ROI. Although the selected patches over the whole brain can provide strong information about the structure of images, they are not as discriminative as patches extracted in some key regions such as hippocampus. In addition, many subjects in the ADNI dataset have vascular lesions (Cuingnet et al., 2011), which may result in some noisy patches when they are extracted over the whole brain. Therefore, extracting patches in a set of key regions instead of the whole brain may provide an improved classification accuracy, which is also indicated in Cuingnet et al. (2011). In the end, the proposed mi-Graph using the whole brain can achieve classification accuracies of 0.877 for CN vs AD and 0.674 for SMCI vs PMCI using a leave-one-out cross validation.

3.3. Comparison with linear SVM

The performance of the proposed mi-Graph was compared with the performance of the widely used linear SVM. In order to use the standard linear SVM, the K patches in a bag were arranged to form a single feature vector. Then, the feature vectors and the corresponding image labels were used as input for the linear SVM. The cost parameter C was determined using a grid search for both the proposed mi-Graph and the linear SVM. To ensure a fair comparison, the number of selected patches and the spatial distance threshold were optimized for both methods. Table 2 shows a comparison of the classification performances between these two methods. The classification accuracy (ACC), sensitivity (SEN), specificity (SPE), positive predictive value (PPV) and negative predictive value (NPV) are presented in the table. The p-values of McNemar tests are also shown in Table 2 to assess the performances of these two methods. In addition, receiver operating characteristic (ROC) curves of these two methods for different classification tasks are given in Fig. 7. As we can see from Table 2 and Fig. 7, the proposed method achieves a significantly more accurate performance than the linear SVM ($p < 0.05$). The improvement is gained by replacing linear kernels with graph kernels in SVM. Although the linear SVM classification uses the same patch selection step, it just utilizes the intensities within patches and neglects the relationships among these patches, resulting in a lower classification accuracy than the proposed mi-Graph.

3.4. Comparison with state-of-the-art methods

A comprehensive study (Cuingnet et al., 2011) compared ten methods using 509 baseline MR images, which reported good

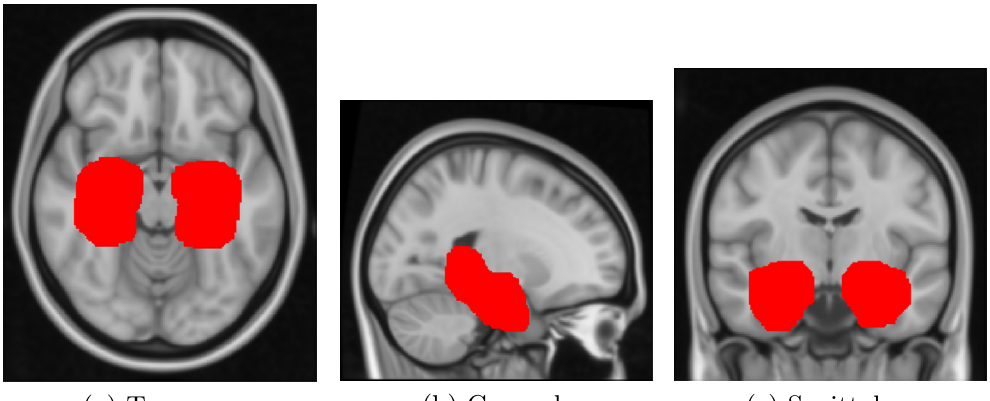


Fig. 4. Orthogonal views of the ROI around hippocampus in the MNI152 space used to select discriminative patches.

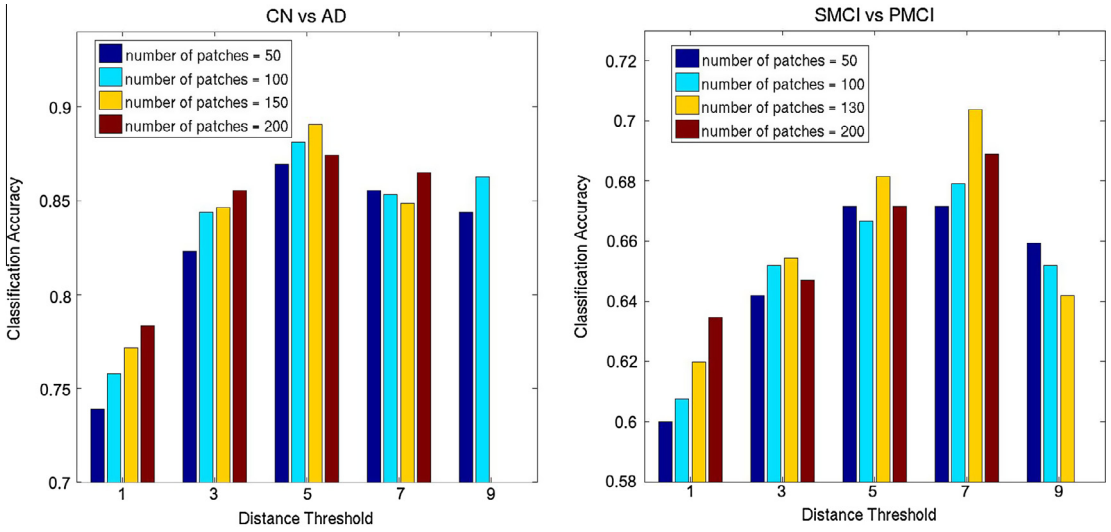


Fig. 5. Effect of the spatial patch distance threshold and the number of patches on the classification accuracy.

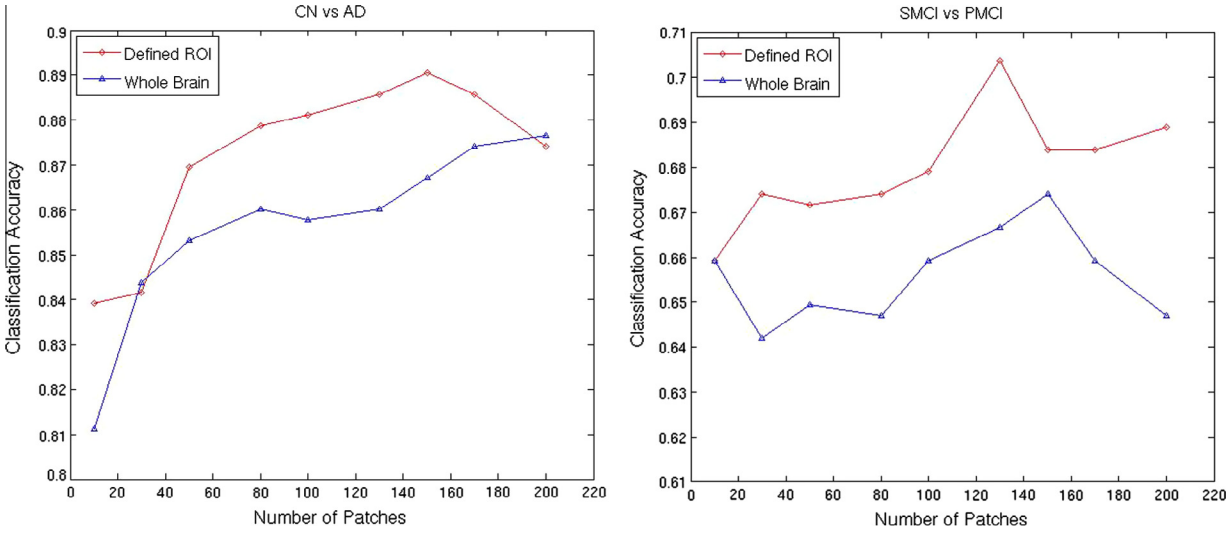


Fig. 6. Comparison between the results when selecting patches over the whole brain and in a predefined ROI around hippocampus. The spatial distance thresholds were set to 4 voxels for CN vs AD and 7 voxels for PMCI vs SMCI when patches were selected over the whole brain. They were set to 5 and 7 voxels for CN vs AD and PMCI vs SMCI respectively when patches were selected within the defined ROI.

Table 2
Method comparison. The number of selected patches and the spatial distance threshold were determined through cross validation and the optimal parameter settings were chosen for both methods. The *p*-value of the McNemar test is shown to assess the performance of the mi-Graph in comparison with the linear SVM.

Comparison	Method	ACC (%)	SEN (%)	SPE (%)	PPV (%)	NPV (%)
CN vs AD ($p = 0.0096$)	Linear SVM	86.0	78.8	92.2	89.7	83.5
	mi-Graph	89.0	84.9	92.6	90.8	87.7
CN vs PMCI ($p = 0.0081$)	Linear SVM	79.4	61.1	92.6	85.7	76.7
	mi-Graph	82.9	68.9	93.1	87.8	80.5
PMCI vs SMCI ($p = 0.0382$)	Linear SVM	66.4	58.7	71.9	59.4	71.3
	mi-Graph	70.4	66.5	73.1	63.4	75.6

results in the classification of CN vs AD. However, it also shows that only four methods can discriminate SMCI and PMCI slightly more accurately than a random classifier (Cuingnet et al., 2011). In a recent study (Wolz et al., 2011), a multi-method was proposed to combine different imaging biomarkers and obtained more accurate results than all the ten methods compared in Cuingnet et al. (2011) on the same dataset. In this paper, we compared our proposed mi-Graph with the multi-method proposed in Wolz et al. (2011). All the 834 baseline scans in the ADNI database were used for evaluation and a leave 5% out cross validation was adopted as in Wolz et al. (2011). 5% of the dataset was randomly selected as the testing set and the remaining 95% was treated as the training set. This process was repeated 100 times and the average results are reported. Studies shows that it is difficult to make a fair comparison of methods if not exactly the same data and cross validation approaches are used (Wolz et al., 2011). Therefore, we performed

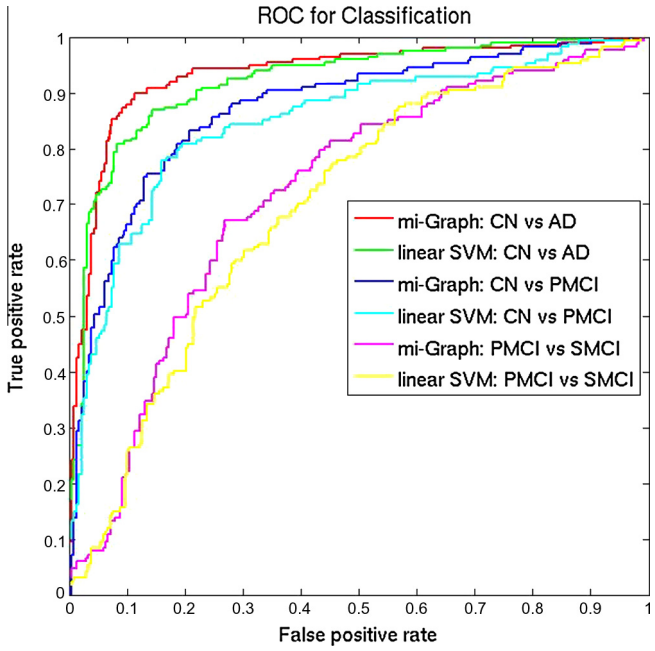


Fig. 7. ROC curves of different methods for the classification of CN vs AD, CN vs PMCI and SMCI vs PMCI.

our experiments on the same 834 images and utilized the same cross validation for a fair comparison. Table 3 shows the results using the proposed mi-Graph and the multi-method. The results using the multi-method are different from the best results reported in Wolz et al. (2011) because a different evaluation scheme was used in Wolz et al. (2011). For CN vs AD, mi-Graph achieved more accurate classification results than methods using hippocampal (HC) volume (Lötjönen et al., 2011), cortical thickness (Lerch and Evans, 2005), manifold-based learning (Wolz et al., 2012) and tensor-based morphometry (Koikkalainen et al., 2011). When all the four types of imaging biomarkers were combined, the multi-method obtained a classification accuracy of 87%. In the proposed method, an accuracy of 90% can be achieved for the classification of CN vs AD. The classification between SMCI and PMCI is far more challenging because the anatomical changes at the prodromal stage of AD disease are subtle (Coupé et al., 2012). For PMCI vs SMCI, the classification accuracies obtained by mi-Graph are higher than all the accuracies presented in Wolz et al. (2011), which demonstrates the effectiveness of the proposed method.

Table 3
Comparison of classification results between mi-Graph and the multi-method proposed in Wolz et al. (2011). Results are obtained using leave 5% out cross validation and this process was repeated 100 times.

Methods	Features	CN vs AD ACC–SEN–SPE	PMCI vs SMCI ACC–SEN–SPE
Multi-Method (Wolz et al., 2011)	Hippocampal volume	80–81–79%*	66–65–67%*
	Cortical thickness	83–71–92%*	47–47–47%*
	Manifold-based learning	85–82–87%*	68–71–66%
	Tensor-based morphometry	85–78–91%*	65–61–68%*
	All	87–78–95%*	67–69–66%*
mi-Graph	Intensity patches	90–86–93%	72–69–74%

* Statistically significant different from the proposed mi-Graph method with $p < 0.001$.

Table 4
Comparison of classification results between mi-Graph and SNIPE proposed in Coupé et al. (2012). Results are obtained using leave-one-out cross validation. HC and EC represent hippocampus and entorhinal cortex.

Methods	Features	CN vs AD ACC–SEN–SPE	PMCI vs SMCI ACC–SEN–SPE
SNIPE (Coupé et al., 2012)	HC volume + Age	79–76–82%	62–61–63%
	HC grade + Age	88–83–92%	71–70–71%
	HC grade + Volume + Age	87–83–91%	71–70–72%
	HC-EC volume + Age	78–76–80%	63–63–64%
	HC-EC grade + Age	89–84–93%	70-69–71%
mi-Graph	Intensity patches	89–85–93%	70–67–73%

A recent study (Coupé et al., 2012) also utilized the same 834 images for evaluation and represents the best published results on the classification of AD when only the ADNI baseline scans are used. They presented various results using different cross validation approaches. For 10-fold cross validation or half training and half testing validation (Cuingnet et al., 2011), the classification accuracies have a high variation when different samplings of the dataset are used (Coupé et al., 2012). Since the leave-one-out cross validation is known to be an almost unbiased estimator (Cawley and Talbot, 2004), we used this validation to evaluate our method and compared the performance with that of the SNIPE (Scoring by Nonlocal Image Patch Estimator) method proposed in Coupé et al. (2012). The advantage of SNIPE is that it utilizes a large amount of non-local patches, which could provide a rich information for the disease label propagation. By contrast, we only use a small subset of relevant patches in our work and investigates the relationships among these training patches for classification. Table 4 shows the comparison between SNIPE and mi-Graph. As we can see, mi-Graph obtained similar accuracies as those when grading values and ages were used as features in SNIPE, which are the best results presented in Coupé et al. (2012) when a leave-one-out cross validation was used.

4. Discussion

In this paper, we have developed a patch-based approach for the classification of subjects with disease such as AD. Since patches that are extracted from images of patients with AD may not be affected by AD or affected by other types of diseases (i.e. cerebrovascular disease), there is some ambiguity in assigning disease labels to these patches. We proposed to use MIL to address the problem of ambiguous labels of the training patches. The intensities of the patches and the relationships among these patches were integrated into graphs and then utilized for classification. We have found that the normalized intensities can provide enough discriminative information for classification and can yield promising results if a good feature selection method is used. By using a simple linear SVM, classification accuracies of 86% for CN vs AD and 66% for SMCI vs PMCI can be achieved. The linear SVM classifier just uses the appearance information of patches (nodes) in the graphs and considers each patch as local isolated measure, ignoring the relationship between patches. However, not only the nodes but also the connections between nodes are affected by the atrophy of AD. Therefore, it is a more effective way to capture the subtle changes by using both the information of nodes and connections in the graphs. When using both the intensities of patches and the relationships among them, the proposed mi-Graph can achieve classification accuracies of 89% for CN vs AD and 70% for SMCI vs PMCI using leave-one-out cross validation, which shows significant improvement over the linear SVM. These results demonstrate that the relationships among patches can provide complementary and useful information for classification.

The effect of the parameters in the patch selection step was also analyzed in our study. When varying the patch selection parameters, including the number of patches and the spatial patch distance threshold, the attributes of the constructed graphs are affected. The aim of the patch extraction step is to find well-connected patches and then establish meaningful graphs which can capture the inherent structural information of the images and discriminate between groups. The optimization of these parameters is crucial for the construction of discriminative graphs for classification. A better search of the K meaningful patches may yield an improved classification accuracy. It should also be mentioned that some patches extracted from healthy subjects may be affected by other types of pathologies such as cerebrovascular disease. This will affect the attributes of the constructed graphs corresponding to these healthy subjects and may degrade the classification performance of the proposed method. However, our proposed method can still achieve promising classification performance even though noisy patches may be present, which also demonstrates the robustness of the proposed mi-Graph method.

Recent studies reported classification accuracies in the range of 76–94% in identifying AD over CN and 64–82% in classifying PMCI over SMCI (Wolz et al., 2011). The high variation in these reported results is partly due to different study populations being used but also due to the fact that different studies using different validation approaches, e.g. leave-one-out or 10-fold cross-validation. In addition, not all reported studies avoid the double dipping bias, e.g. by using the combined training and test data in the feature selection stage (Eskildsen et al., 2013). In this paper, the locations of selected patches were determined without using the test data to avoid the double-dipping bias. We also compared the performance of the proposed method with those of two recent papers (Wolz et al., 2011; Coupé et al., 2012) which are among the most competitive approaches and represent state-of-the-art using ADNI MR baseline images. We evaluated the proposed method on the same dataset and used the same cross validation approaches as in these two papers. This allows direct comparisons with these state-of-the-art methods. The results demonstrate the effectiveness of the proposed method and indicate that the proposed method could provide another potential choice for the early detection of AD. Furthermore, we also evaluated the proposed mi-Graph method on the standardized set of baseline scans from ADNI-1 (Wyman et al., 2013) to enable future comparisons with other methods. The classification results using these 818 baseline MR scans at 1.5 T are shown in Table 5.

In another recent publication (Liu et al., 2013), a hierarchical ensemble classification method also utilized patches as features and derived correlative features among these patches to improve the classification accuracy. Then, multiple local classifiers were hierarchically fused to build a strong classifier. This method demonstrates that an improved classification can be achieved by combining the patches and the correlative features. This method can achieve an accuracy of 92% on the classification of AD over CN using a ten fold cross validation. However, the authors do not report the classification accuracy for SMCI vs PMCI and hence we cannot make a direct comparison in this scenario. The major difference between mi-Graph and this ensemble classification method is that in their paper the authors used a hierarchical approach to fuse different classifiers while only a global classifier was trained in our proposed method. In addition, the authors used gray matter densities for extracting discriminative patches, whereas intensities were directly used in our work, therefore eliminating the requirement for any segmentation of the images.

Table 5
Classification results of the proposed mi-Graph method using the standardized ADNI-1 dataset, which consists of 818 baseline MR scans at 1.5 T.

Comparison	ACC	SEN	SPE	PPV	NPV
CN vs AD	89.2%	85.1%	92.6%	90.4%	88.3%
PMCI vs SMCI	69.3%	66.7%	71.2%	62.6%	71.2%

Although the proposed method can yield competitive accuracies compared with state-of-the-art methods, it also has some limitations. First, the intensities of patches were used as features, which may be susceptible to distortion artefacts, intensity imhomogeneity, or other forms of noise. Therefore, the preprocessing steps such as image denoising or intensity normalization will affect the final classification results. More optimized preprocessing steps could improve the performance of the proposed mi-Graph method. Second, the same distance thresholds were used for all patches in the patch selection process in our work. It may be possible to improve the classification performance of the proposed method if an autocorrelation metric could be used to search the optimal distance thresholds for different patches. In addition, it should be noted that the graph kernel used in this paper may not be the best choice. A better graph kernel may be able to capture more useful structural information of the images. For example, a graph kernel which can combine the information of nodes, edges, and topological structures for comparison may improve the classification performance. Especially using topological information, which is not considered in this paper, may provide complementary information for our classification work. Graph kernels based on subgraphs (Huan et al., 2005) or graphlets (Shervashidze et al., 2009) could be used to capture this information. Finally, non-rigid registrations were used to align all the images. This step removes inter-subject anatomical variations but may also remove some of the pathological changes. We believe that there is a trade-off between the level of non-rigid alignment which can ensure that corresponding brain structures are well aligned and the amount of pathological changes which can still allow us to measure subject-specific differences for classification. We will investigate the effect of using different levels of registrations on the classification performance of the proposed method in future work.

Other interesting work could also be investigated in future. For example, in our work, the trained classifier is a bag-level classifier, which can only predict the label of unseen images and cannot assign a probabilistic label for every voxel. It would be clinically more useful if the method enables the visualization of differences between groups (Coupé et al., 2012). Therefore, in future work we will try to learn instance-level classifiers (Andrews et al., 2002) to assign a disease score for every voxel. This may provide more clinical insight of the pathological changes of AD. In addition, it is reported that the correct diagnosis rate in the ADNI study is estimated at 90% (Ranginwala et al., 2008). Due to the uncertainty of diagnosis, individual subjects might be wrongly categorized. Our proposed method can address the problem of ambiguous labels at the patch level but not at the image level. It would be interesting to examine the weakly supervised method to overcome the problem of ambiguous training labels of ADNI images due to misdiagnosis. Moreover, the proposed method has been evaluated on the baseline MR scans in the ADNI database. 3D spatial graphs were built for each subject, which contain the information about the structures of the images at single time point. Using the longitudinal datasets, this approach could be extended to 4D graphs and these constructed graphs may then capture more complementary information about the morphological changes over time. We believe that graphs based on space and time will capture more discriminative information and yield an improved classification.

5. Conclusion

In this study, we have shown that the multiple instance learning technique can be successfully applied to the classification of AD. The proposed method was evaluated on a large database using the entire 834 baseline MR scans in the ADNI study. The direct comparisons with two recent methods demonstrate the effectiveness of the proposed method. In future work, we plan to extend the proposed framework using longitudinal datasets and other imaging modalities, such as FDG-PET images.

Acknowledgments

This project was partially funded by the China Scholarship Council. 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.

References

Amari, S.-i., Wu, S., 1999. Improving support vector machine classifiers by modifying kernel functions. Neural Networks 12 (6), 783–789.

Andrews, S., Tsochantaridis, I., Hofmann, T., 2002. Support vector machines for multiple-instance learning. Adv. Neural Inf. Process. Syst. 15, 561–568.

Babenko, B., Yang, M.-H., Belongie, S., 2009. Visual tracking with online multiple instance learning. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 983–990.

Bi, J., Liang, J., 2007. Multiple instance learning of pulmonary embolism detection with geodesic distance along vascular structure. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8.

Brookmeyer, R., Johnson, E., Ziegler-Graham, K., Arrighi, H.M., 2007. Forecasting the global burden of Alzheimers disease. Alzheimer's Dementia 3 (3), 186–191.

Cawley, G.C., Talbot, N.L., 2004. Fast exact leave-one-out cross-validation of sparse least-squares support vector machines. Neural Networks 17 (10), 1467–1476.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2, 27:1–27:27, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Cho, Y., Seong, J.-K., Jeong, Y., Shin, S.Y., 2012. Individual subject classification for Alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data. Neuroimage 59 (3), 2217–2230.

Chu, C., Hsu, A.-L., Chou, K.-H., Bandettini, P., Lin, C.-P., 2011. Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. NeuroImage 60, 59–70.

Clark, C., Davatzikos, C., Borthakur, A., Newberg, A., Leight, S., Lee, V.-Y., Trojanowski, J., 2007. Biomarkers for early detection of Alzheimer pathology. Neurosignals 16 (1), 11–18.

Coupé, P., Eskildsen, S.F., Manjón, J.V., Fonov, V.S., Pruessner, J.C., Allard, M., Collins, D.L., 2012. Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease. NeuroImage: Clinical 1 (1), 141–152.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., Chupin, M., Benali, H., Colliot, O., 2011. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. Neuroimage 56 (2), 766–781.

Eskildsen, S.F., Coupé, P., García-Lorenzo, D., Fonov, V., Pruessner, J.C., Collins, D.L., 2013. Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. NeuroImage 65, 511–521.

Fan, Y., Shen, D., Gur, R.C., Gur, R.E., Davatzikos, C., 2007. COMPARE: classification of morphological patterns using adaptive regional elements. IEEE Trans. Med. Imaging 26 (1), 93–105.

Fu, Z., Robles-Kelly, A., Zhou, J., 2011. MILIS: multiple instance learning with instance selection. IEEE Trans. Pattern Anal. Mach. Intell. 33 (5), 958–977.

Gerardin, E., Chételat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.-S., Niethammer, M., Dubois, B., Lehéricy, S., Garnero, L., et al., 2009. Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. Neuroimage 47 (4), 1476.

Ghosh, D., Chinnaiyan, A.M., 2005. Classification and selection of biomarkers in genomic data using LASSO. BioMed Res. Int. 2005 (2), 147–154.

Gray, K.R., Wolz, R., Heckemann, R.A., Aljabar, P., Hammers, A., Rueckert, D., 2012. Multi-region analysis of longitudinal FDG-PET for the classification of Alzheimer's disease. NeuroImage 60 (1), 221–229.

Gray, K.R., Aljabar, P., Heckemann, R.A., Hammers, A., Rueckert, D., 2013. Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. Neuroimage 65, 167–175.

Grundman, M., Petersen, R.C., Ferris, S.H., Thomas, R.G., Aisen, P.S., Bennett, D.A., Foster, N.L., Jack Jr., C.R., Galasko, D.R., Doody, R., et al., 2004. Mild cognitive impairment can be distinguished from Alzheimer disease and normal aging for clinical trials. Arch. Neurol. 61 (1), 59.

Herholz, K., Salmon, E., Perani, D., Baron, J., Holthoff, V., Frölich, L., Schönknecht, P., Ito, K., Mielke, R., Kalbe, E., et al., 2002. Discrimination between Alzheimer dementia and controls by automated analysis of multicenter FDG-PET. Neuroimage 17 (1), 302–316.

Hinrichs, C., Singh, V., Mukherjee, L., Xu, G., Chung, M.K., Johnson, S.C., 2009. Spatially augmented lpboosting for AD classification with evaluations on the ADNI dataset. Neuroimage 48 (1), 138–149.

Hinrichs, C., Singh, V., Xu, G., Johnson, S.C., 2011. Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. Neuroimage 55 (2), 574–589.

Huan, J., Bandyopadhyay, D., Wang, W., Snoeyink, J., Prins, J., Tropsha, A., 2005. Comparing graph representations of protein structure for mining family-specific residue-based packing motifs. J. Comput. Biol. 12 (6), 657–671.

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

Jack Jr, C.R., Knopman, D.S., Jagust, W.J., Petersen, R.C., Weiner, M.W., Aisen, P.S., Shaw, L.M., Vemuri, P., Wiste, H.J., Weigand, S.D., et al., 2013. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. Lancet Neurol. 12 (2), 207–216.

Janousova, E., Vounou, M., Wolz, R., Gray, K., Rueckert, D., Montana, G., 2012. Biomarker discovery for sparse classification of brain images in Alzheimer's disease. Ann. Brit. Mach. Vision Assoc.(BMVA) 2012 (2), 1–11.

Keihaninejad, S., Zhang, H., Ryan, N.S., Malone, I.B., Modat, M., Cardoso, M.J., Cash, D., Fox, N.C., Ourselin, S., 2013. An unbiased longitudinal analysis framework for tracking white matter changes using diffusion tensor imaging with application to Alzheimer's disease. NeuroImage 72, 153–163.

Koikkalainen, J., Lötjönen, J., Thurfjell, L., Rueckert, D., Waldemar, G., Soininen, H., 2011. Multi-template tensor-based morphometry: application to analysis of Alzheimer's disease. NeuroImage 56 (3), 1134–1144.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S., Baker, C.I., 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci. 12 (5), 535–540.

Lerch, J.P., Evans, A.C., 2005. Cortical thickness analysis examined through power analysis and a population simulation. Neuroimage 24 (1), 163–173.

Leung, K.K., Barnes, J., Modat, M., Ridgway, G.R., Bartlett, J.W., Fox, N.C., Ourselin, S., 2011. Brain MAPS: an automated, accurate and robust brain extraction technique using a template library. Neuroimage 55 (3), 1091–1108.

Liu, M., Zhang, D., Shen, D., 2012a. Ensemble sparse classification of Alzheimer's disease. Neuroimage 60 (2), 1106–1116.

Liu, M., Zhang, D., Yap, P.-T., Shen, D., 2012b. Tree-Guided Sparse Coding for Brain Disease Classification, MICCAI, pp. 239–247.

Liu, M., Zhang, D., Shen, D., 2013. Hierarchical fusion of features and classifier decisions for Alzheimer's disease diagnosis. Human Brain Mapping doi: 10.1002/hbm.22254.

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

Lu, L., Bi, J., Wolf, M., Salganicoff, M., 2011. Effective 3D object detection and regression using probabilistic segmentation features in CT images. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1049–1056.

Maron, O., Lozano-Pérez, T., 1998. A framework for multiple-instance learning. Adv. Neural Inform. Process. Syst., 570–576.

Mika, S., Ratsch, G., Weston, J., Scholkopf, B., Mullers, K., 1999. Fisher discriminant analysis with kernels. In: IEEE Workshop on Neural Networks for Signal Processing IX. IEEE, pp. 41–48.

Neuhaus, M., Bunke, H., 2007. A quadratic programming approach to the graph edit distance problem. IAPR Workshop on Graph-Based Representations in Pattern Recognition, 92–102.

Nyúl, L.G., Udupa, J.K., 1999. On standardizing the MR image intensity scale. Magn. Reson. Med. 42 (6), 1072.

Pihlajamäki, M., Sperling, R.A., 2008. fMRI: use in early Alzheimers disease and in clinical trials. Future Neurol. 3 (4), 409–421.

Pless, R., Souvenir, R., 2009. A survey of manifold learning for images. IPSJ Trans. Comput. Vision Applic. 1 (0), 83–94.
Proverb, C., 1995. A probabilistic atlas of the human brain: theory and rationale for its development. Neuroimage 2, 89–101.
Ranginwala, N.A., Hynan, L.S., Weiner, M.F., White III, C.L., 2008. Clinical criteria for the diagnosis of Alzheimer disease: still good after all these years. Am. J. Geriatric Psych 16 (5), 384–388.
Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J., 1999. Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans. Med. Imaging 18 (8), 712–721.
Sanchez, A., David, V., 2003. Advanced support vector machines and kernel methods. Neurocomputing 55 (1), 5–20.
Schölkopf, B., Smola, A., Müller, K.-R., 1997. Kernel principal component analysis. In: International Conference on Artificial Neural Networks. Springer, pp. 583–588.
Shervashidze, N., Petri, T., Mehlhorn, K., Borgwardt, K.M., Viswanathan, S., 2009. Efficient graphlet kernels for large graph comparison. In: International Conference on Artificial Intelligence and Statistics, 2009, pp. 488–495.
Tong, T., Wolz, R., Gao, Q., Hajnal, J.V., Rueckert, D., 2013. Multiple Instance Learning for Classification of Dementia in Brain MRI, MICCAI, pp. 599–606.
Vounou, M., Janousova, E., Wolz, R., Stein, J.L., Thompson, P.M., Rueckert, D., Montana, G., 2012. Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. Neuroimage 60 (1), 700–716.
Wee, C.-Y., Yap, P.-T., Li, W., Denny, K., Browndyke, J.N., Potter, G.G., Welsh-Bohmer, K.A., Wang, L., Shen, D., 2011. Enriched white matter connectivity networks for accurate identification of MCI patients. Neuroimage 54 (3), 1812–1822.
Wee, C.-Y., Yap, P.-T., Shen, D., 2012a. Prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns. Human Brain Mapping doi: 10.1002/hbm.22156.
Wee, C.-Y., Yap, P.-T., Zhang, D., Denny, K., Browndyke, J.N., Potter, G.G., Welsh-Bohmer, K.A., Wang, L., Shen, D., 2012b. Identification of MCI individuals using structural and functional connectivity networks. Neuroimage 59 (3), 2045–2056.
Wolz, R., Heckemann, R.A., Aljabar, P., Hajnal, J.V., Hammers, A., Lötjönen, J., Rueckert, D., et al., 2010. Measurement of hippocampal atrophy using 4D graph-cut segmentation: application to ADNI. NeuroImage 52 (1), 109.
Wolz, R., Julkunen, V., Koikkalainen, J., Niskanen, E., Zhang, D.P., Rueckert, D., Soininen, H., Lötjönen, J., 2011. Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease. PloS one 6 (10), e25446.
Wolz, R., Aljabar, P., Hajnal, J.V., Lötjönen, J., Rueckert, D., 2012. Nonlinear dimensionality reduction combining MR imaging with non-imaging information. Medical Image Anal. 16 (4), 819–830.
Wyman, B.T., Harvey, D.J., Crawford, K., Bernstein, M.A., Carmichael, O., Cole, P.E., Crane, P.K., DeCarli, C., Fox, N.C., Gunter, J.L., et al., 2013. Standardization of analysis sets for reporting results from ADNI MRI data. Alzheimer's Dementia 9 (3), 332–337.
Xu, Y., Zhang, J., Chang, E., Lai, M., Tu, Z., 2012. Context-constrained multiple instance learning for histopathology image segmentation, MICCAI, pp. 623–630.
Yiannopoulou, K.G., Papageorgiou, S.G., 2013. Current and future treatments for Alzheimers disease. Therap. Adv. Neurol. Disorders 6 (1), 19–33.
Yoon, U., Lee, J.-M., Im, K., Shin, Y.-W., Cho, B.H., Kim, I.Y., Kwon, J.S., Kim, S.I., 2007. Pattern classification using principal components of cortical thickness and its discriminative pattern in schizophrenia. Neuroimage 34 (4), 1405–1415.
Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., 2011. Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55 (3), 856–867.
Zhou, Z.-H., Sun, Y.-Y., Li, Y.-F., 2009. Multi-instance learning by treating instances as non-IID samples. Int. Conf. Mach. Learn., 1249–1256.