NeuroImage 59 (2012) 2217–2230

Contents lists available at SciVerse ScienceDirect

NeuroImage

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

Individual subject classification for Alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data

Youngsang Cho [a], Joon-Kyung Seong [b,*], Yong Jeong [c,d], Sung Yong Shin [a] and for the Alzheimer's Disease Neuroimaging Initiative [1]

[a] Computer Science Department, KAIST, Republic of Korea
[b] School of Computer Science and Engineering, Soongsil University, Republic of Korea
[c] Department of Bio and Brain Engineering, KAIST, Republic of Korea
[d] Department of Neurology, Samsung Medical Center, Republic of Korea

ARTICLE INFO

Article history:
Received 28 April 2011
Revised 22 September 2011
Accepted 29 September 2011
Available online 8 October 2011

Keywords:
Individual subject classification
Alzheimer's disease
Cortical thickness
Frequency representation
Incremental learning

ABSTRACT

Patterns of brain atrophy measured by magnetic resonance structural imaging have been utilized as significant biomarkers for diagnosis of Alzheimer's disease (AD). However, brain atrophy is variable across patients and is non-specific for AD in general. Thus, automatic methods for AD classification require a large number of structural data due to complex and variable patterns of brain atrophy. In this paper, we propose an incremental method for AD classification using cortical thickness data. We represent the cortical thickness data of a subject in terms of their spatial frequency components, employing the manifold harmonic transform. The basis functions for this transform are obtained from the eigenfunctions of the Laplace–Beltrami operator, which are dependent only on the geometry of a cortical surface but not on the cortical thickness defined on it. This facilitates individual subject classification based on incremental learning. In general, methods based on region-wise features poorly reflect the detailed spatial variation of cortical thickness, and those based on vertex-wise features are sensitive to noise. Adopting a vertex-wise cortical thickness representation, our method can still achieve robustness to noise by filtering out high frequency components of the cortical thickness data while reflecting their spatial variation. This compromise leads to high accuracy in AD classification. We utilized MR volumes provided by Alzheimer's Disease Neuroimaging Initiative (ADNI) to validate the performance of the method. Our method discriminated AD patients from Healthy Control (HC) subjects with 82% sensitivity and 93% specificity. It also discriminated Mild Cognitive Impairment (MCI) patients, who converted to AD within 18 months, from non-converted MCI subjects with 63% sensitivity and 76% specificity. Moreover, it showed that the entorhinal cortex was the most discriminative region for classification, which is consistent with previous pathological findings. In comparison with other classification methods, our method demonstrated high classification performance in both categories, which supports the discriminative power of our method in both AD diagnosis and AD prediction.

Introduction

Objectives

Alzheimer's disease (AD) is the most common form of dementia. The incidence of AD doubles every five years after age of 65 (Bain et al., 2008). As life expectancy increases, the number of AD patients increases accordingly, which causes a heavy socioeconomic burden. Currently used treatments offer a small symptomatic benefit in mild to moderate AD but cannot delay or halt the progression of AD. Recently, Jack et al. (2010) reported that a structural change was observed in human brains a few years before any symptomatic awareness. Thus, the structural change could provide a clue for early detection of AD.

Amnestic mild cognitive impairment (MCI) is known as a prodromal state of AD and has received much attention for early detection of AD. While only 2% of the healthy elders were converted to AD every year, 10–15% of amnestic MCI patients were converted to AD annually (Petersen et al., 1999). However, AD cannot be predicted with only MCI diagnosis, because other forms of dementia can also be preceded by the MCI state (Dubois and Albert, 2004). Some MCI patients were converted to AD and others remained stable or even reversed to a normal status. The former is called MCI converters (MCIc) and the latter is MCI non-converters (MCInc). The classification between MCInc and MCIc was dealt with for early detection of AD (Misra et al., 2009;

* Corresponding author. Fax: +82 2 822 3622..
E-mail address: joon.swallow@gmail.com (J.-K. Seong).

[1] Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (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. ADNI investigators include (complete listing available at: http://adni.loni.ucla.edu/wp-content/uploads/how to apply/ADNI Authorship List.pdf.)

1053-8119/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2011.09.085

Querbes et al., 2009). Since AD is known to reduce brain volumes prior to clinical symptoms of dementia, brain atrophy has also been utilized for this purpose.

However, the classification between MCInc and MCIc is challenging. According to the benchmark results in Cuingnet et al. (2011), classification methods in this category showed low accuracies. Cuingnet et al. (2011) stated that low accuracies can be caused certainly by heterogeneity of cortical thinning patterns over MCInc subjects. One can overcome the difficulty in two ways: increasing training data so as to cover all complex patterns, and choosing good features so as to reflect the difference between two groups.

A large volume of data is necessary in order to handle complex patterns of data. Moreover, simultaneous acquisition of such a data set is practically difficult. Because new data are obtained usually from physical examination or during disease diagnosis, the volume of data is increasing steadily. The continuing data acquisition makes conventional classification methods ineffective. Whenever a new data set is obtained, a classification method should train its classifier with an entire data set including the new data set in order to additively reflect the latter of which the volume is small in general. In this study, we employed an incremental learning approach in order to address this issue effectively.

For accurate classification, it is essential to choose eligible features which clearly represent group differences. Subcortical structures such as hippocampi and certain regions of gray matter were substantially more vulnerable in AD; especially, they showed atrophy in an early stage of AD. Accordingly, most of AD classification methods based on brain morphometry utilized one of the following three neuroanatomical features: hippocampal features, tissue probability maps, and cortical thickness data. Among these features, we used cortical thickness data rather than hippocampal features and probability maps. Hippocampal features were known to be difficult to capture exactly (Khan et al., 2008; Qiu and Miller, 2008). The previous study by Cuingnet et al. (2011) also showed that classification methods using hippocampal features were less accurate compared to other methods. The methods based on tissue probability maps are dependent on volume registration, while the methods based on cortical thickness data are dependent on surface registration. In general, surface registration provided better correspondences than volume registration (Anticevic et al., 2008; Desai et al., 2005). Therefore, we adopt cortical thickness data for AD classification.

Recently, several classification methods based on cortical thickness data have been proposed. Desikan et al. (2009) and Querbes et al. (2009) classified HC, MCI, and AD using the mean values of cortical thickness for neuroanatomically segmented regions as a feature vector. However, the region-wise data cannot reflect local characteristics of smaller regions than the segmented ones. Cuingnet et al. (2011) constructed a classifier using the cortical thickness at every vertex as a feature vector. Although the vertex-wise data can reflect local deformity of a small region, they are sensitive to noise and registration errors. In this study, we overcome the difficulties with both types of features by adopting the noise-filtered vertex-wise cortical thickness data based on spatial frequency analysis

We present an individual subject classification method based on incremental learning for AD diagnosis and AD prediction using the cortical thickness data. These data are mapped onto a spatial frequency domain from the surface of a cortex by using the manifold harmonic transform. The basis functions for this transformation are smooth and periodic with different frequencies. Since high frequency components are sensitive to noise and registration errors rather than group differences, we cut off these components to filter out noise, which also effectively reduces the dimension of data as observed in lossy data compression. We construct our classifier based on principal component analysis (PCA) and linear discriminant analysis (LDA) (Zhao et al., 2003), which enables incremental learning for AD diagnosis and AD prediction. The efficacy of the proposed classification method was demonstrated using several experiments, including comparison with ten other well-known classification methods and also validation of incremental classification.

Previous work

Numerous studies on brain morphometry in AD have been conducted in the past two decades. It turns out that AD tends to deform brain regions. For example, volume reduction of temporal lobes was observed in brains of AD patients, according to studies based on voxel-based morphometry (Chételat et al., 2005; Good et al., 2002; Karas et al., 2003). Reduction of cortical thickness was also observed on cortical surfaces, in particular, medial temporal lobes and entorhinal cortices (Dickerson et al., 2009; Lerch et al., 2005; Thompson et al., 2003). In (Wang et al., 2006), volume reduction and shape deformity of hippocampi were shown in mild AD. Although these studies reported new findings on AD, they did not provide automatic tools for AD diagnosis and AD prediction. However, these findings were significant in studies on AD classification since vulnerable structures to AD can be excellent features for these tasks.

Based on the results of such morphometry studies, many classification methods have been proposed for AD diagnosis and AD prediction recently. Colliot et al. (2008) and Gerardin et al. (2009) utilized hippocampal volumes and hippocampal shapes for AD classification, respectively. Colliot et al. (2008) normalized hippocampal volumes with respect to the total intracranial volume, and the average of the two normalized hippocampal volumes for both hemispheres was used to classify HC, MCI and AD. In Gerardin et al. (2009), hippocampal shapes were modeled via the spherical harmonics transform, and then the support vector machine (SVM) was employed to find a hyperplane to maximally separate HC and AD. A technical limitation of these methods was that extracted subcortical structures from MR images were often unacceptable due to their small shapes and ill-defined boundaries (Chupin et al., 2007). Klöppel et al. (2008) applied the SVM to voxels of a tissue probability map which indicates the probability of different tissue classes (gray matter, white matter, and cerebrospinal fluid). The approach can be adapted easily to other anatomical features due to its methodological simplicity. Specifically, Cuingnet et al. (2011) applied it to cortical thickness data for AD classification. However, the approach was sensitive to noise and registration errors because the spatial coherence of features was not considered. Unlike the method in Klöppel et al. (2008), other methods parcellated MR volumes or cortical surfaces, and extracted a representative value such as the average cortical thickness from each region. Ye et al. (2008) and Magnin et al. (2009) classified AD and HC by using the tissue probability of each parcellated region. In Desikan et al. (2009); Querbes et al. (2009), the cortical thickness value of each segmented region was extracted to build a feature vector. Ye et al. (2008), Desikan et al. (2009), Magnin et al. (2009), and Querbes et al. (2009) parcellated MR volumes based on neuroanatomical knowledge, and Fan et al. (2007) adaptively parcellated MR volumes to define discriminative regions. However, based on region segmentations, such methods poorly reflected detailed spatial variation of features. Furthermore, none of the above methods addressed the incremental learning issue in classification.

To the best of our knowledge, incremental learning has not been attempted in neuroimaging. However, since most of training data for classification are sequentially obtained, studies on incremental learning have received steady attentions in artificial intelligence and computer vision. Incremental learning-based versions of statistical techniques such as PCA and LDA have been reported. Hall et al. (1998) proposed an incremental PCA method which updates the PCA transformation matrix sequentially with each of the additional training data. Hall et al. (2002) extended it to deal with a set of new training data simultaneously. Levy and Lindenbaum (2000) and Lim et al. (2004) later enhanced the computational efficiency and reduced

the computational error during the update. The method of Lim et al. (2004) was used in a visual tracking method (Ross et al., 2008) in order to adapt to changes in the appearance of the target. Pang et al. (2005) proposed an incremental LDA (ILDA) scheme which can handle large training data. They showed that the classification accuracy increased incrementally with additional data.

Materials

Data used in this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). 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), non-profit organizations and private pharmaceutical companies. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (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 in an early stage of AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as to reduce the time and cost of clinical trials. For more details, we refer the readers to www.adni-info.org.

In this paper, MRIs of 491 subjects who belong to one of HC, MCI, and AD groups were analyzed. The eligibility criteria of subjects applied in ADNI are described at http://www.adni-info.org/Scientists/ADNIGrant/ProtocolSummary.aspx. Briefly, enrolled subjects in ADNI were between the ages of 55 and 90 (inclusive) and spoke either English or Spanish. All subjects must be willing and able to undergo all test procedures including neuroimaging and agree to longitudinal follow-up. Specific psychoactive medications were excluded. General inclusion/exclusion criteria are as follows:

1. HC subjects: Mini-Mental State Examination (MMSE) (Folstein et al., 1975) scores between 24 and 30 (inclusive), a Clinical Dementia Rating (CDR) of 0, non-depressed, non-MCI, and nondemented. The age range of normal subjects was roughly matched to that of MCI and AD subjects. Therefore, there should be minimal enrollment of normals under the age of 70.
2. MCI subjects: MMSE scores between 24 and 30 (inclusive), a memory complaint, objective memory loss measured by education-adjusted scores on Wechsler Memory Scale Logical Memory II, a CDR of 0.5, absence of significant levels of impairment in other cognitive domains, essentially preserved activities of daily living, and an absence of dementia.
3. Mild AD subjects: MMSE scores between 20 and 26 (inclusive), CDR of 0.5 or 1.0, and meeting NINCDS/ADRDA criteria for probable AD.

In the ADNI procedure, all subjects received the baseline clinical/cognitive examinations including 1.5 T structural MRI, and were re-evaluated at specified intervals (6 or 12 months) for 2–3 years. The baseline MRI scans were downloaded from the ADNI database, which were used as the input data in our experiments. We also utilized the follow-up examination results in order to separate MCI into MCIc and MCInc. The separation was performed by following the same policy as that in Cuingnet et al. (2011). MCI subjects who were converted to AD within 18 months were classified into MCIc, and those not converted into AD within the same period were classified into MCInc. Further, we excluded MCI subjects who did not follow up the examinations more than 18 months. Table 1 shows the demographic characteristics of the participants.

Table 1
Demographic characteristics of HC, MCInc, MCIc, and AD.

	Age, years	Sex, M/F	MMSE score
HC (n = 160)	76.2 ± 5.4 (60–90)	74/86	29.2 ± 1.0 (25–30)
MCInc (n = 131)	74.1 ± 7.2 (58–88)	81/50	27.2 ± 1.7 (24–30)
MCIc (n = 72)	74.8 ± 7.6 (55–88)	41/31	26.5 ± 1.8 (23–30)
AD (n = 128)	76.0 ± 7.1 (55–91)	60/68	23.3 ± 2.0 (20–27)

n: the number of subjects in a group.
Data for age, years of education and MMSE score: mean ± SD (range).
Data for sex: the number of subjects.
M:male, F:female.

Incremental classification method

Overview

In this section, we present an incremental classification method for AD diagnosis and AD prediction using cortical thickness data. Given an MR volume of a subject, AD diagnosis determines whether she/he is in HC or AD. On the other hand, AD prediction discriminates MCIc from MCInc, provided with an MR volume of an MCI subject. We also deal with classification between HC and MCIc for comparison with the benchmark results in Cuingnet et al. (2011). Our classification method consists of two steps: group classifier training and individual subject classification. Fig. 1 shows an overall structure of the method. The former step trains a group classifier with labeled MR volumes. This step first filters out high frequency components from the cortical thickness data at vertices, which has been extracted from the MR volumes, in order to remove noise, and then trains the group classifier with resulting data. The latter step classifies unlabeled subjects by using an individual subject classifier. This classifier is initialized with the group classifier trained in the previous step and incrementally updated. Given the MR volume of a subject, its feature vector representing the noise-filtered cortical thickness data is acquired as in group classifier training. The classifier performs AD diagnosis or AD prediction using the feature vector. Unlike existing methods for AD classification, our individual subject classifier not only classifies an unlabeled subject but also enhances the classifier itself based on incremental learning after labeling the subject.

The contributions of our approach are two-fold: first, we represent the cortical thickness data of a subject in terms of their frequency components, employing the manifold harmonic transform (Levy, 2006; Qiu et al., 2006; Vallet and Lévy, 2008). The basis functions for this transform are obtained from the eigenvectors of the Laplace–Beltrami (LB) operator, which is dependent only on the geometry of a cortical surface but not on the cortical thickness function defined on it. This facilitates individual subject classification based on incremental learning. Second, our classifier showed high accuracy in both AD diagnosis and AD prediction. According to the benchmark data in Cuingnet et al. (2011), no AD classifiers produced good results in the both categories. Intuitively, methods based on region-wise features poorly reflect the detailed spatial variation of cortical thickness, and those based on vertex-wise features are sensitive to noise. Adopting a vertex-wise cortical thickness representation scheme, we can still achieve robust classification to noise by filtering out high frequency components of cortical thickness data while reflecting their spatial variation. We believe that this compromise enabled both AD diagnosis and AD prediction with high accuracy.

Group classifier training

Feature vector construction

We describe how to construct a feature vector from an MR volume. A feature vector should reflect group differences as much as possible so as to achieve high classification performance. It should also be compact so as to achieve computational efficiency. We use the cortical thickness data at the vertices of the cortical surface to

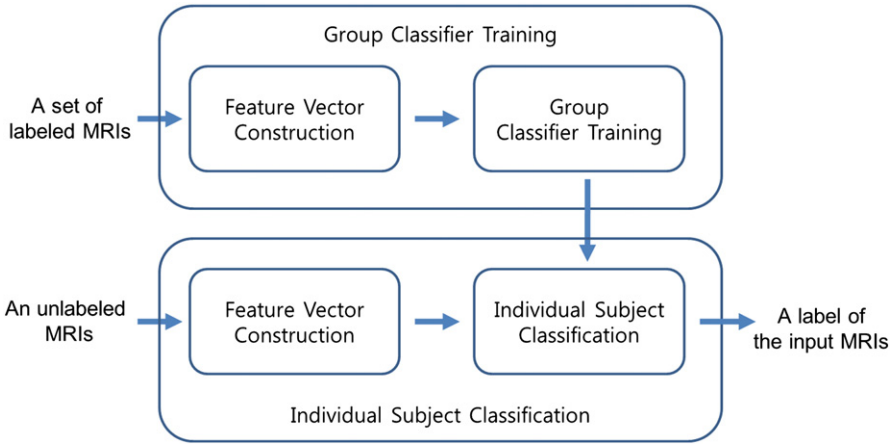

Fig. 1. Overview of the proposed classification method.

construct a feature vector based on the observation that the cortex of an AD subject becomes thinner in specific regions at an early stage of AD progression (Querbes et al., 2009). Our feature vector construction scheme consists of two steps: cortical thickness data extraction and noise removal. Although the two-step scheme is separately applied to both hemispheres, we describe the construction scheme for only one hemisphere for explanation simplicity.

Given an MR volume for a subject, we extract cortical thickness data from it by using version 4.4.0 of the FreeSurfer software package (Athinoula A. Martinos Center at the Massachusetts General Hospital, Harvard Medical School; http://www.surfer.nmr.mgh.harvard.edu/), which is a popular free software for cortical surface analysis. We first construct the outer and inner cortical surface meshes from the MR volume. As the outer mesh is generated by deforming the inner mesh, the two meshes are isomorphic, that is, have the same number of vertices and the same connectivity. Therefore, their vertices with the same index correspond to each other. The cortical thickness at each vertex is defined as the distance between the two surfaces at the vertex (Singh et al., 2008). However, in order to establish the correspondences between the vertices of different subjects, the cortical thickness data of each subject needs to be represented in a common space. We use the atlas surface provided by the FreeSurfer software as the common space, and register each cortical surface to the atlas surface using this software. In general, the FreeSurfer software does not enforce isomorphic meshes for all subjects. However, we can make the meshes isomorphic by remeshing them after registration using an executable file, mri_surf2surf in the software. The atlas surface is represented with a mesh $M=\{V,E\}$, where $V=\{v_i|i=1,\ldots,n\}$ and $E=\{e_{ij}=(v_i,v_j)|i<j$ and $j\leq n\}$ are the vertex set and the edge set, respectively. We approximate the cortical thickness function $C(s)$ on the atlas surface S with the cortical thickness vector $c=(c_1,\ldots,c_n)^T$, where $c_i=C(v_i)$, $1\leq i\leq n$ is the cortical thickness at vertex v_i.

Given a cortical thickness vector c of a subject, we now describe how to remove noise. Our noise removal scheme maps $C(s)$ from the surface S onto a frequency domain, and then discards high frequency components. Accordingly, the dimension of the cortical thickness data is reduced. Chung et al. (2007) and Shen et al. (2007) used the spherical harmonic transform (SHT) to map the cortical thickness and hippocampal shape of a subject onto frequency domains, respectively. However, their methods require mappings from 3D surfaces to a sphere as preprocessing which causes an inherent distortion because of shape differences. Recently, a scheme called the manifold harmonic transform (MHT) has been introduced for spectral analysis of scalar functions defined on a surface (Levy, 2006; Qiu et al., 2006; Vallet and Lévy, 2008). The MHT represents such a scalar function in terms of its frequency components by using a set of basis functions. Specifically, the eigenfunctions of the Laplace–Beltrami (LB) operator are adopted as the basis functions of the MHT. Levy (2006) introduced the MHT into computer graphics and discussed its possible applications, and Vallet and Lévy (2008) used the MHT to edit 3D models in the frequency domain. In neuroimaging, the MHT has been employed for smoothing neuroanatomical features defined on cortical surfaces (Qiu et al., 2006; Seo et al., 2010; Kim et al., accepted for publication; Seo and Chung, 2011). Seo et al. (2011) also proposed to extract the centerline of the segmented human mandible based on the second LB eigenfunction. Unlike the SHT, the MHT maps a scalar function from a 3D surface onto a frequency domain without employing an extra mapping to a sphere (Levy, 2006; Qiu et al., 2006; Vallet and Lévy, 2008). Recently, Seo and Chung (2011) showed that the MHT is superior to the SHT in representing neuroanatomical functions defined on cortical surfaces: The MHT can achieve a more precise representation with a less number of basis functions. For our purpose, we need an additional mapping from cortical surfaces to an atlas surface. A spherical mapping results in surface flattening which is rarely observed in a non-spherical mapping. We adopt the MHT rather than the SHT to avoid the distortion due to surface flattening.

Provided with $C(s)$, the MHT represents it with its frequency components $(f_1,\ldots,f_n)^T$ where $f_i, 1\leq i\leq n$ are defined as follows:

$$f_k = C, H^k = \int_S C(s)H^k(s)ds \tag{1}$$

$$\approx c, h^k = \sum_{i=1}^{n} A_i c_i h_i^k, k = 1,\ldots,n. \tag{2}$$

Here $H^k(s), k=1,\ldots,n$ is the eigenfunction of the continuous LB operator of which the corresponding eigenvalue is the kth smallest one among all eigenvalues, and is approximated by the kth eigenvector $h^k=(h_1^k,\ldots,h_n^k)^T$ of the discrete LB operator on the mesh M, where A_i is one third of the area sum of triangles sharing vertex v_i. Given the frequency components $(f_1,\ldots,f_n)^T$, the cortical thickness data $c=(c_1,\ldots,c_n)^T$ can be reconstructed as follows:

$$c = \sum_{k=1}^{n} f_k h^k. \tag{3}$$

We adopt the method proposed by Vallet and Lévy (2008) to compute the eigenfunctions. For details in eigenfunction computation, we refer the readers to Vallet and Lévy (2008).

Each eigenfunction corresponds to a specific frequency which is equal to the square root of its corresponding eigenvalue (Vallet and Lévy, 2008). Fig. 2 visualizes eigenvectors on the left cortical surface where h^k denotes the kth eigenvector. The value of the kth

eigenvector for large k changes periodically with a short cycle. Therefore, such an eigenfunction corresponds to a high frequency component of $C(s)$.

Since high frequency components of cortical thickness data mainly represent noise and individual variabilities irrelevant to the characteristics of a group, classification performance is less sensitive to high frequency components than low frequency ones. Chung et al. (2007) showed that the power of statistical analysis increases by reducing the intensities of high frequency components. Inspired by this result, we remove noise and thus reduce the dimension of cortical thickness data by filtering out the high frequency components f_i, $i > F$, where F is the cut-off dimension for the cortical thickness data. The noise-filtered data of a subject is represented by a reduced vector $(f_1,\ldots,f_F)^T$. Employing the scheme in Qiu et al. (2008), the cut-off dimension F is determined by goodness of fit G, that is,

$$G = \frac{\sum_{j=1}^{N} \left\| c^j - \sum_{k=1}^{F} f_k^j h^k \right\|^2}{\sum_{j=1}^{N} \left\| c^j \right\|^2}, \tag{4}$$

where c^j is the original cortical thickness data of subject j in a training set, f_k^j is the kth frequency component of subject j, and N is the number of subjects. In the numerator, $\sum_{k=1}^{F} f_k^j h^k$ represents the noise-filtered cortical thickness data of subject j for the cut-off dimension F. Qiu et al. (2008) set $G = 0.05$ to determine the cut-off dimension F. For our experiments, we conservatively set $G = 0.025$, which is one half of the value used by them. With this value of G, the cut-off dimension F was fixed to 2400. Given the cut-off dimension F for the (reduced) vectors for both hemispheres of a brain, we define a feature vector $x = (f_1^l,\ldots, f_F^l, f_1^r,\ldots,f_F^r)^T$, where f_i^l and f_i^r, $i = 1,2,\ldots,F$ denote the ith frequency components for the left and right hemispheres, respectively.

In PCA, the principal components are recomputed whenever a training data set changes. However, the LB eigenfunctions are dependent only on the shape of a surface (Levy, 2006; Qiu et al., 2006), which is common to all scalar functions defined on it under the assumption that the template surface is fixed over all subjects. Therefore, the cortical thickness should be measured with the same template; specifically, we used the one available in the FreeSurfer software. Unlike PCA, the LB operator requires no retraining even if different scalar functions on the same surface are used.

Group classifier training

In this section, we explain how to train the group classifier with the feature vectors obtained from a set of labeled subjects. Our group classifier is based on PCA and LDA (Belhumeur et al., 1997; Liu and Wechsler, 2000; Yu and Yang, 2001). We could improve classification performance with more sophisticated statistical techniques such as PLS (Partial Least Squares) (Liu and Rayens, 2007) and OPLS (Oriented Partial Least Squares) (Bylesjø et al., 2006). However, we employed the most basic classification scheme, the PCA–LDA method (Belhumeur et al., 1997) for clarity of presentation and easy extension to incremental learning. LDA is a statistical technique that finds coordinate axes which maximally separate different groups of data. This technique has been widely used for applications such as face recognition and speech recognition. Although LDA showed high classification accuracy in many applications (Pang et al., 2005), it suffered from the singularity of scatter matrices when dealing with a small number of training samples and a high dimensional feature space. The combination of PCA and LDA resolves this problem by reducing the dimension of feature vectors with PCA.

Given feature vectors x_i, $1 \le i \le N$ which belong to one of g groups $\{G_1,\ldots,G_g\}$, our group classifier is trained by performing PCA and LDA in sequence. In order to perform PCA, we first derive the covariance matrix V of the training data set $X = \{x_1,\ldots,x_N\}$:

$$V = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T \tag{5}$$

where $\bar{x}$ is the mean of all feature vectors. Since each x_i is represented as a feature vector $(f_1^l,\ldots,f_F^l,f_1^r,\ldots,f_F^r)^T$, V is an $2F \times 2F$ matrix. In our classification problem, the covariance matrix V is singular in general, which results in the singularity problem in performing LDA. To prevent this problem, we choose the eigenvectors of V corresponding to the largest k non-zero eigenvalues to construct a PCA transformation matrix W_P. The number of chosen eigenvectors determines the dimension of the reduced space (PCA space). A popular heuristic for deciding the dimension k is based on the percentage of the total variance achieved with the largest k eigenvalues. In Jolliffe (2002), it is suggested that the percentage between 70% and 90% preserves most of information needed for representing a data set with a Gaussian-like distribution. We empirically decided the dimension k by setting the percentage to 70%, which worked well for our experiments.

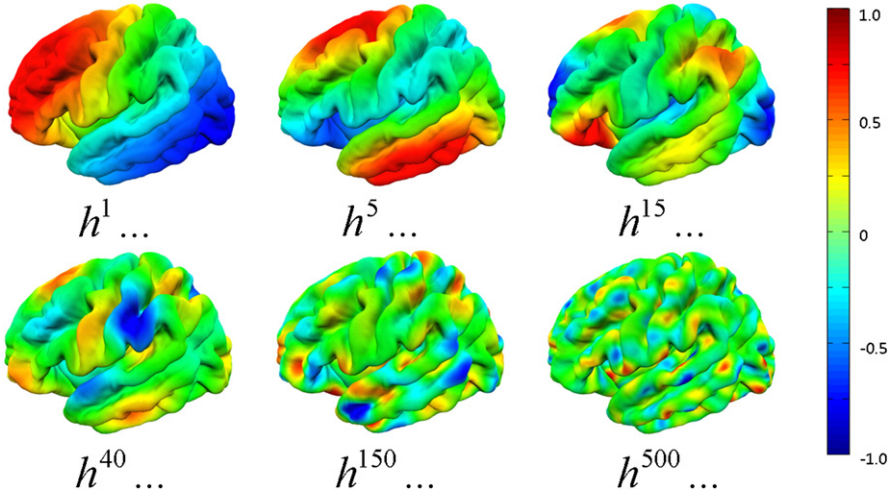

Fig. 2. Visualization of eigenvectors for the LB operator on the left cortical surface mesh: h^k denotes the kth eigenfunction. The value of the kth eigenvector for large k changes periodically with a short cycle on the mesh.

More sophisticated schemes for deciding k are available in Zhu (2006). Given the PCA transformation matrix W_P, a feature vector x in the feature space is converted to a vector y in a PCA space spanned by the column vectors of W_P as follows:

$$y = W_P^T x. \tag{6}$$

By applying PCA to all feature vectors in X, we obtain a new training set $Y = \{y_i | y_i = W_P^T x_i, i = 1,\ldots,N\}$ in the PCA space.

We then conduct LDA with the training data set Y. LDA finds the coordinate axes which maximally separate the groups of the data set. In Fig. 3, the axis w_1 for a data set better separates two groups than the axis w_2. The mean difference between the groups is larger on w_1 than on w_2. On the other hand, the variance within each group is smaller on w_1 than on w_2. LDA maximizes the between-classes variance of Y across the groups and minimizes the within-class variance for each group. Therefore, LDA finds an axis w that maximizes the following energy function (Balakrishnama and Ganapathiraju, 1998):

$$J(w) = \frac{\sigma_{between}(w)}{\sigma_{within}(w)} = \frac{w^T S_B w}{w^T S_W w}, \tag{7}$$

where $\sigma_{between}(w)$ and $\sigma_{within}(w)$ are the between-classes variance and the within-class variance projected onto the axis w, respectively. The between-classes scatter matrix S_B and the within-class scatter matrix S_W are defined as follows:

$$S_B = \sum_{i=1}^{g} N_i (\bar{y}_i - \bar{y})(\bar{y}_i - \bar{y})^T, \text{ and} \tag{8}$$

$$S_W = \sum_{i=1}^{g} \sum_{k=1}^{N_i} (y_{ik} - \bar{y}_i)(y_{ik} - \bar{y}_i)^T = \sum_{i=1}^{g} N_i V_i, \tag{9}$$

where $\bar{y}_i$, V_i and N_i are the mean, the covariance matrix and the size of group i, respectively, and y_{ik} is the kth feature vector in the group G_i. Notice that the within-class scatter matrix S_W can be invertible by choosing the PCA space dimension as described previously. The problem of finding the axis maximizing J is reduced to an eigenvalue problem $S_W^{-1} S_B w = \lambda w$ by differentiating Eq. (7) with respect to w (Fisher, 1936). Therefore, the optimal axis is the eigenvector of the matrix $S_W^{-1} S_B$ with the largest eigenvalue (see Appendix A for detailed derivation). The LDA coordinate system has the axes specified by the eigenvectors with a small number of the largest eigenvalues. Therefore, an input data y from the PCA subspace is mapped onto the LDA space as follows:

$$z = W_L^T y, \tag{10}$$

where the matrix W_L is the LDA transformation matrix consisting of the chosen eigenvectors. Since $S_W^{-1} S_B$ has at most $g-1$ eigenvectors with non-zero eigenvalues, the dimension of the LDA space is always less than g.

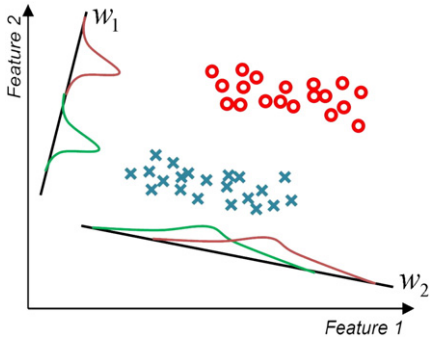

Fig. 3. 2D illustration of LDA classification for two groups: LDA finds the coordinate axes which maximally separate the groups of a data set. In this figure, the axis w_1 for the data set better separates two groups than the axis w_2. Specifically, the mean difference between the groups is larger on w_1 than on w_2, while the variance within each group is smaller on w_1 than on w_2. LDA maximizes the between-classes variance across the groups and minimizes the within-class variance for each group.

Individual subject classification

In this section, we describe how to classify individual subjects using the group classifier trained in the previous section. We also discuss how to enhance the performance of the classifier incrementally with new training data. Given an MR volume of an unlabeled subject, its feature vector is first acquired as described in Feature vector construction section. This process performs PCA and LDA in sequence to transform the feature vector x of the subject to a point z in the LDA space:

$$z = W_L^T W_P^T x, \tag{11}$$

where W_P and W_L are the PCA and LDA transformation matrices defined in Group classifier training section, respectively. Given z in the LDA space, our classifier maps the subject onto one of the groups $\{G_1, G_2, \ldots, G_g\}$. That is, the subject is classified to group G_i if z is closer to the projection of the mean $\bar{x}_i$ of G_i onto the LDA space than those of the others. After the unlabeled subject is classified, its label is validated by a clinician to be used for training. It is time-consuming to train the classifier with the entire training data whenever new training data are added. Therefore, the group classifier training scheme is slightly modified for training an individual subject classifier. In other words, the individual subject classifier is trained incrementally without using the previously used training data so that the time efficiency is dependent only on the size of new data.

Suppose that the group classifier has been built with a training data set, $X_1 = \{x_1, \ldots, x_N\}$. The classifier can be modeled with a set of parameters, $\Omega = \{W_P, W_L, \bar{x}_1, \ldots, \bar{x}_g\}$, where $\bar{x}_i$, $1 \le i \le g$ is the mean of feature vectors x_{ik}, $1 \le k \le N_i$ belonging to group G_i. Given anadditional data set, $X_2 = \{x_{N+1}, \ldots, x_{N+M}\}$ for small $M \ge 1$, our objective is to update Ω incrementally with X_2 in order to obtain a new model $\Omega' = \{W_P, W_L, \bar{x}_1, \ldots, \bar{x}_g\}$.

In order to update W_P, we first modify the mean vector $\bar{x}$ and the covariance matrix V with X_2, and then derive the new PCA matrix W_P' from these, by employing an incremental PCA method (Lim et al., 2004). For clarity of explanation, suppose that $M = 1$, that is, $X_2 = \{x_{N+1}\}$ is a singleton set. The new mean $\bar{x}$ is computed from the old mean $\bar{x}$ as follows:

$$\bar{x} = \frac{(N\bar{x} + x_{N+1})}{(N+1)}. \tag{12}$$

The new covariance matrix V' is

$$V' = \frac{N}{N+1} V + \frac{N}{(N+1)^2} (x_{N+1} - \bar{x})(x_{N+1} - \bar{x})^T. \tag{13}$$

For details in deriving the transformation matrix W_P, we refer the readers to Hall et al. (1998). Hall et al. (2002) proposed an incremental PCA method which updates a classifier with a set of feature vectors rather than a single feature vector, and Lim et al. (2004) enhanced the computational efficiency for incremental PCA. We adopt the method in Lim et al. (2004), the source code of which is available at http://www.cs.toronto.edu/ dross/code/.

With W_p, $\bar{x}$, $\bar{x}_1, \ldots, \bar{x}_g$ computed, we employ an incremental LDA method (Pang et al., 2005) to compute the LDA transformation matrix W_L. Unlike the group classifier described in Group classifier training section, this method incrementally computes the between- and within-class scatter matrices in the feature space (rather than in the PCA subspace) and then maps the results onto the PCA subspace to finally compute W_L. Let S_B^X and S_W^X be the between- and within-class scatter matrices in the feature space, respectively. These matrices

can be computed with a small number of parameters including group means $\bar{x}, \bar{x}_1, \ldots, \bar{x}_g$ and group covariance matrices $V_1, \ldots, V_g$ in the feature space, which can efficiently be updated in an incremental manner. We describe how to convert S_W^X in the feature space to S_W in the PCA subspace using the PCA transformation matrix W_P:

$$\begin{aligned} S_W &= \sum_{i=1}^{g}\sum_{k=1}^{N_i}(y_{ik}-\bar{y}_i)(y_{ik}-\bar{y}_i)^T \\ &= \sum_{i=1}^{g}\sum_{k=1}^{N_i}\left(W_P^T x_{ik}-W_P^T\bar{x}_i\right)\left(W_P^T x_{ik}-W_P^T\bar{x}_i\right)^T \\ &= \sum_{i=1}^{g}\sum_{k=1}^{N_i}W_P^T(x_{ik}-\bar{x}_i)(x_{ik}-\bar{x}_i)^T W_P \\ &= W_P^T\left(\sum_{i=1}^{g}\sum_{k=1}^{N_i}(x_{ik}-\bar{x}_i)(x_{ik}-\bar{x}_i)^T\right)W_P \\ &= W_P^T S_W^X W_P, \end{aligned}$$

where x_{ik} is the kth feature vector in the group G_i. Similarly, S_B^X can be converted to S_B, that is, $S_B = W_P^T S_B^X W_P$. Finally, W_L are constructed with eigenvectors of $S_B^{-1}S_W$.

Results

The proposed method was validated through three experiments: noise removal, group classification performance, and incremental classification performance. We describe these experiments in the following three sections, respectively. For all experiments, we used the entire data summarized in Table 1. These data were divided into two sets: a training data set and a test data set. For comparison purpose, we used the same training and test data sets as those in Cuingnet et al. (2011). The training data set consists of 80 HC, 65 MCInc, 37 MCIc, and 62 AD, and the test data set has 80 HC, 66 MCInc, 35 MCIc, and 66 AD. For assessment of classification performance, we also prepared three data subsets, Datasets 1, 2, and 3 for HC vs. AD classification, MCInc vs. MCIc classification, and HC vs. MCIc classification, respectively. The training data of each data subset were used for obtaining its corresponding classifier, and the test data of the subset was for validating the classifier. The composition of the two data sets and three data subsets are summarized in Table 2.

In all experiments, we used the cortical thickness data defined on the atlas meshes for the left and right hemispheres each of which consists of 40,962 vertices, that is, the cortical thickness data on each mesh were represented with a 40,962-dimensional vector. The dimension of these data was reduced by noise removal, that is, by cutting off their high frequency components. Except for cortical thickness extraction, the experiments were performed on a PC equipped with an Intel® Core™2 Duo Processor E8500 (3.16 GHz CPU and 12 GB memory). Cortical thickness extraction was done on a cluster computer with sixteen nodes (two Intel Xeon 2.5 GHz CPUs and 32 GB memory for each node). It took about a week to extract cortical thickness data from all MR volumes with the FreeSurfer software. In the remainder of this section, we discuss each experiment in detail.

Noise removal

The first experiment was intended to validate our noise removal scheme described in Feature vector construction section. This scheme removes noise from cortical thickness data by filtering out their high frequency components after determining the cut-off frequency F. We first performed noise removal using the scheme and then performed statistical analysis to compare the cortical thickness data before and after noise removal.

To perform noise filtering, we determined the cut-off dimension F using the training data set consisting of 80 HC, 65 MCInc, 37 MCIc, and 62 AD subjects as given in Table 2. Therefore, the resulting dimension F is unbiased to any specific test data for validation. The cut-off dimension F was chosen by setting the goodness of fit $G = 0.025$. With $G = 0.025$, the cut-off frequency F was set to 2400. Fig. 4 plots accuracy of each of group classifiers with respect to the cut-off dimension F. Notice that we did not try to choose the optimal value of the dimension F in a classification-specific manner as observed in the figure. Rather, we chose a cut-off dimension F for all classifications by conservatively setting $G = 0.025$. This cut-off frequency F was used for all experiments described in Noise removal, Group classification performance and Incremental classification performance sections.

In order to compare cortical thickness data before and after noise removal, we separately performed group analysis for the original cortical thickness and its noise-filtered one using the test data of each data subset in Table 2. Since the group analysis results for all three data subsets were similar to each other, we only present the results for Dataset 1 (HC vs. AD) for conciseness of explanation. We first performed noise removal for the cortical thickness data of each test subject by cutting off their high frequency components using the previously-determined $F = 2400$. We then measured the mean difference between AD and HC groups for the corresponding noise-filtered data subset as well as the original one. The first row in Fig. 5 shows the mean difference in the original data subset, and the second row shows the mean difference in the noise-filtered data subset. One can visually verify that the mean differences are similar to each other. The third row visualizes the difference between the first and second rows: For more than 90% of the whole surface region, the original and noise-filtered cortical thickness data are the same within 0.15 mm error. We also performed a t-test at each vertex in order to validate the hypothesis that the cortical thickness can discriminate AD subjects from HC subjects. The first and the second rows in Fig. 6 show the resulting t-statistics maps for the original and noise-filtered data subsets, respectively. Statistically significant regions in the noise-filtered data subset were similar to those in the original one. The third row shows the difference of absolute t-statistics values at every vertex between the original and noise-filtered data subsets. A warm color represents that the noise-filtered data subset is statistically more significant than the original one, while a cold color represents the opposite case. For more than 63% of the whole surface region, the noise-filtered data subset has greater absolute t-statistics values, which verifies that our noise removal scheme improves the statistical analysis results similarly to the work in Chung et al. (2007).

Group classification performance

The next experiment was to show the group classification performance of our method against other classification methods. Many high-dimensional classification methods for automatic AD classification have recently been presented. Fan et al. (2007) and Querbes et al. (2009) used different data sets from the ADNI database. In order to compare different methods in their classification performances, one would have to implement all these methods. Fortunately, Cuingnet et al. (2011) presented benchmark results for AD classification methods using the ADNI database. The same data from the ADNI database were used to compare the performances of ten classification methods. Since the subject identifications for the data were provided as a supplement, we were able to repeat an identical experiment in Cuingnet et al. (2011) with our method in order to exploit their benchmark results for performance comparison. Now, we briefly explain how the benchmark results were generated.

Cuingnet et al. (2011) performed three classification experiments: HC vs. AD, HC vs. MCIc, and MCIc vs. MCInc. They compared ten

Table 2
Compositions of the two data sets and three data subsets.

	HC	MCInc	MCIc	AD
Training data set	$n = 80$	$n = 65$	$n = 37$	$n = 62$
Test data set	$n = 80$	$n = 66$	$n = 35$	$n = 66$
Dataset 1	✓			✓
Dataset 2		✓	✓	
Dataset 3	✓		✓	

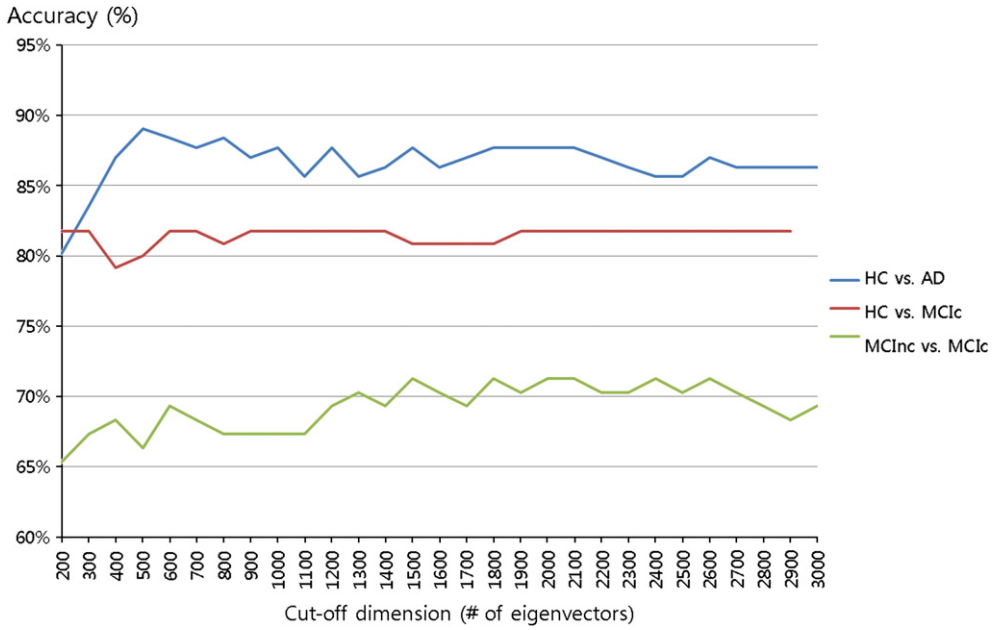

Fig. 4. Accuracies of individual subject classifiers with respect to cut-off dimension F.

classification methods: Voxel-Direct, Voxel-Direct_VOI, Voxel-Atlas, Voxel-STAND, Voxel-COMPARE, Thickness-Direct, Thickness-Atlas, Thickness-ROI, Hippo-Volume and Hippo-Shape. The first five methods employed voxel-based segmented tissue probability maps. The next three methods used cortical thickness data. The last two methods were based on hippocampal features. For details of the classification methods, we refer the readers to Cuingnet et al. (2011). For assessment of classification performances, we used data subsets, Datasets 1, 2, and 3 (see Table 2 for composition of each data subset), which are the same data sets as those in Cuingnet et al. (2011).

We trained three group classifiers for HC vs. AD classification, MCInc vs. MCIc classification, and HC vs. MCIc classification with the training data in their respective data subsets. Specifically, for training a group classifier, we first performed noise removal for the training data in the corresponding data subset using the cut-off frequency $F = 2400$ determined in the previous section. We then employed

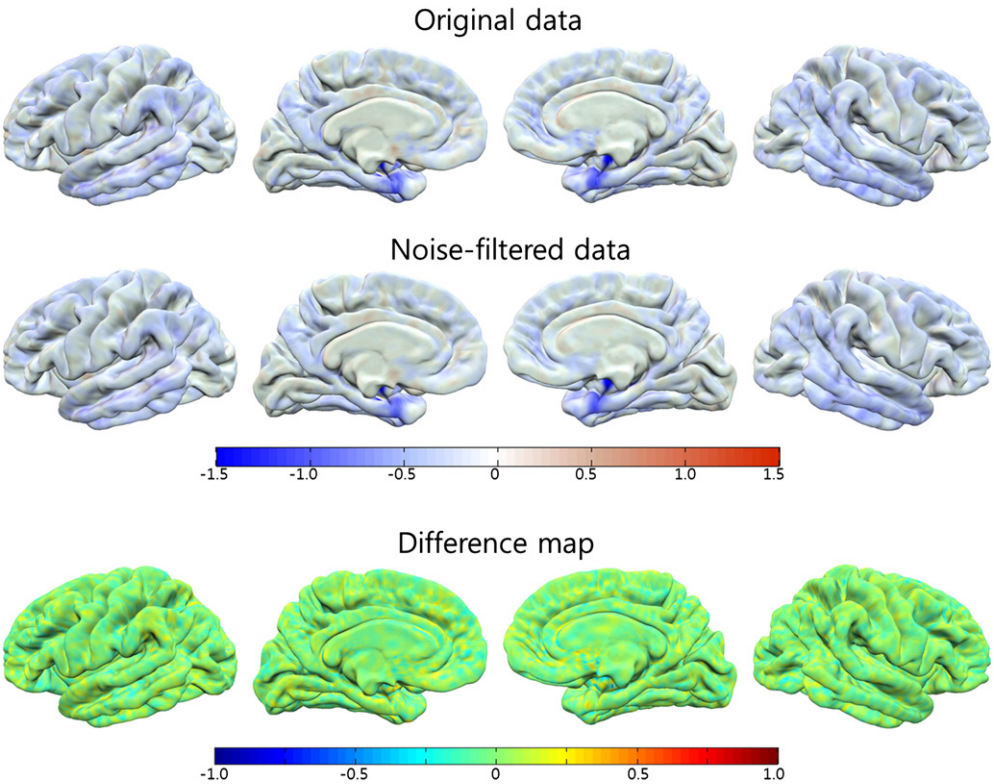

Fig. 5. Visualization of the mean difference of cortical thickness between AD and HC: The first and the second rows show the mean difference in the original data set and the noise-filtered data set, respectively. The third row visualizes the difference between the first and the second rows.

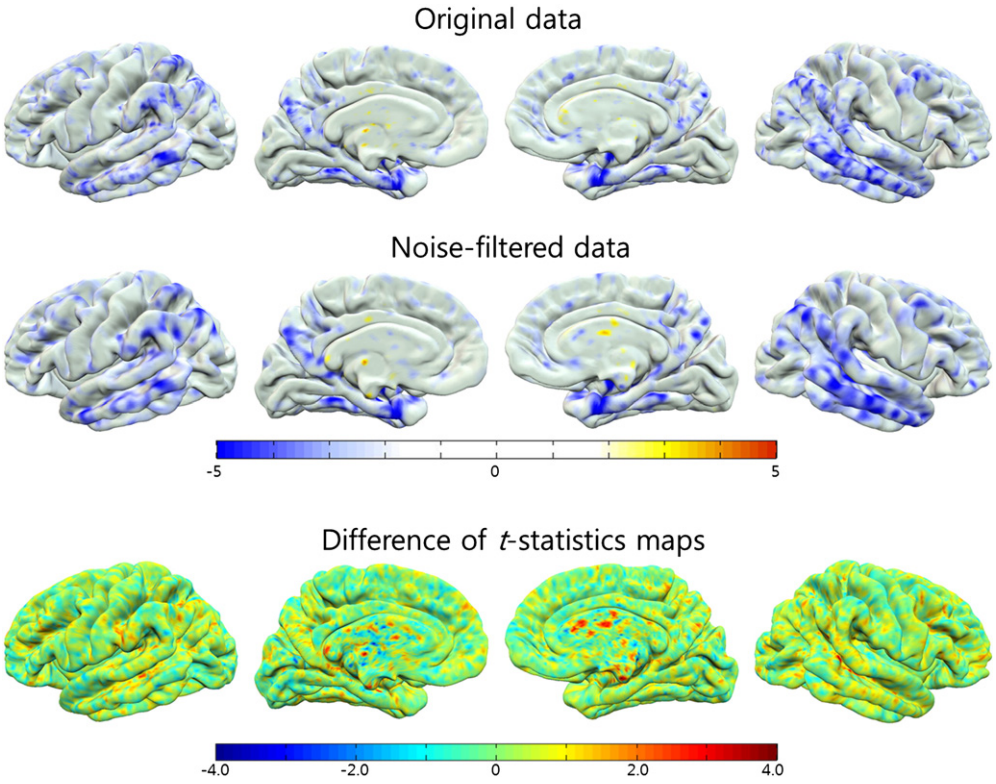

Fig. 6. Comparison of *t*-statistics maps between the original and noise-filtered cortical thickness data: statistically significant regions for the noise-filtered data subset (Dataset 1) were similar to those for the original one. The third row shows the difference of absolute *t*-statistics values at every vertex between the original and noise-filtered data sets. A warm color represents that the noise-filtered data subset is statistically more significant than the original one, while a cold color represents the opposite case. For more than 65% of the whole surface region, the noise-filtered data set has greater *t*-statistics values, which verifies that our noise removal scheme improves the results of statistical analysis.

PCA to reduce the dimension of the feature space, which prevents the singularity problem in performing LDA. As shown in Group classifier training section, we empirically decided this dimension k by setting the percentage of the total variance to 70%. Finally, the group classifier was obtained by performing LDA with the transformed training data in the PCA space.

After training the classifiers, we assessed the sensitivity and the specificity of each classification with the test data in the respective data subset as follows:

$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}.$$

The sensitivity and the specificity were 82% and 93% for HC vs. AD classification, 63% and 76% for MCIc classification, and 66% and 89% for MCIc vs. HC classification, respectively. We compared the classification performance of our method with those of the other ten methods used in Cuingnet et al. (2011). Table 3 summarizes the classification performances of the ten classification methods together with that of ours. The sum of the sensitivity and specificity for each method was used as the measure of classification performance. Fig. 7 depicts the performances of the methods in their descending order. Our method received good evaluations in all classifications: It showed the highest performance in MCInc vs. MCIc classification and the second highest performance in HC vs. AD classification. In HC vs. MCIc classification, it was ranked in the fourth position. In HC vs. AD classification, the performance of ours was similar to that of Voxel-Direct showing the highest performance. In HC vs. MCIc

Table 3
The classification accuracy comparison between the eleven methods.

Methods	HC vs. AD		HC vs. MCIc		MCInc vs. MCIc		Total performance
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	
Ours	82%	93%	66%	89%	63%	76%	469
Voxel-Direct	81%	95%	57%	96%	0%	100%	429
Voxel-Direct-VOI	71%	95%	54%	95%	43%	70%	428
Voxel-STAND	75%	91%	73%	85%	57%	78%	459
Voxel-Atlas	81%	90%	68%	95%	0%	100%	434
Voxel-COMPARE	82%	89%	59%	78%	54%	78%	440
Thickness-Direct	73%	90%	54%	96%	62%	67%	442
Thickness-Atlas	79%	90%	57%	93%	54%	78%	451
Thickness-ROI	69%	94%	65%	94%	24%	82%	428
Hippo-Volume	71%	77%	73%	74%	62%	69%	426
Hippo-Shape	69%	84%	57%	88%	0%	100%	398

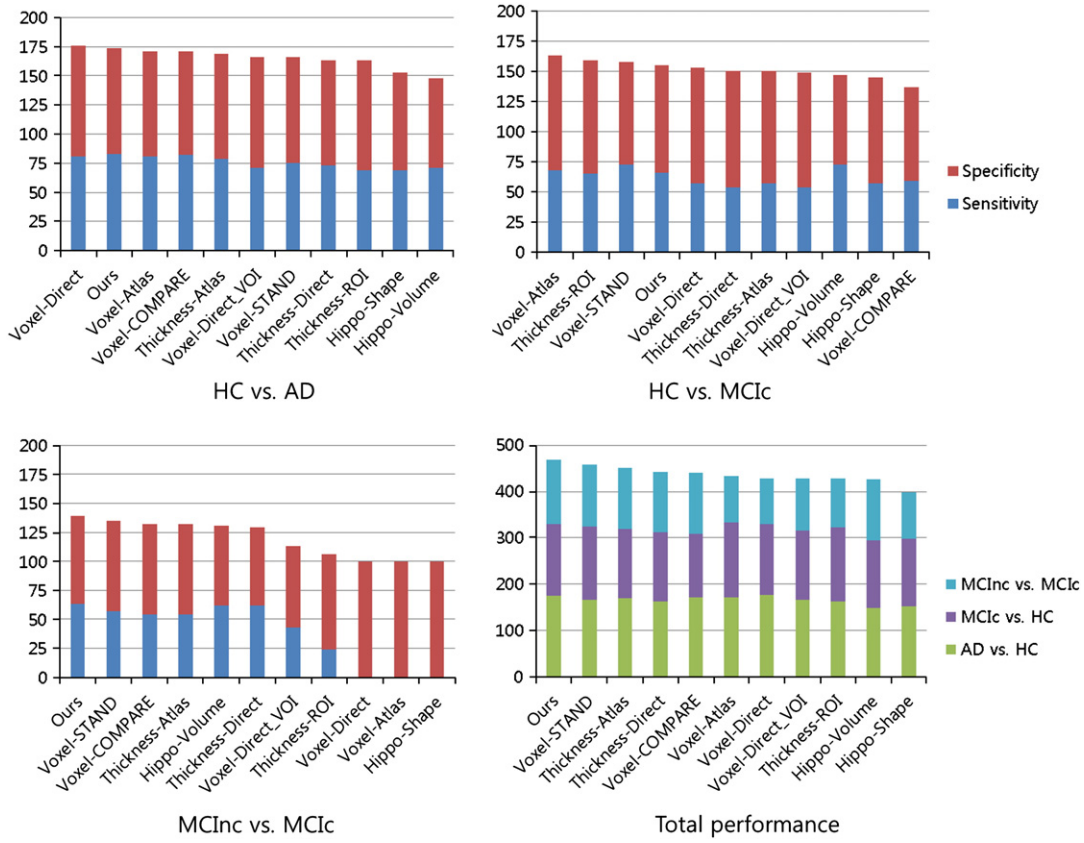

Fig. 7. Benchmark results for eleven classification methods, in which their performances were shown in the descending order. Our method received good evaluations in all classifications: it showed the highest performance in MCInc vs. MCIc classification and the second highest in HC vs. AD classification. In HC vs. MCIc classification, it was ranked in the fourth position. The benchmark results demonstrated that our classification exhibited high performance compared to other classification methods.

classification, three methods, Voxel-Atlas, Thickness-ROI, and Voxel-STAND showed higher performances than ours. However, these four methods showed inferior performances to ours in MCInc and MCIc classifications. Voxel-STAND showed a similar performance to ours in both HC vs. MCIc classification and MCInc and MCIc classifications, but our method is superior to Voxel-STAND in HC vs. AD classification. The benchmark results show that our classification method exhibits its high performance compared to the other methods.

We compared the discriminative regions for our method with those in Thickness-Direct and Thickness-Atlas which used, as feature data, the cortical thickness at every vertex and the mean cortical thickness over every parcellated region, respectively. Both of the methods adopted the support vector machine (SVM) which finds a hyperplane that maximally separates groups. In the case of the linear SVM, the value of the ith component of the vector v that is orthogonal to the separating hyperplane represents the contribution of the component to classification. That is, if the value of the ith component of v is zero, the ith element of every feature vector does not affect the classification result. Conversely, if the value is larger than the others, the classification result is more sensitive to the ith element than the other elements. This vector plays the same role as the separating axis w in LDA since the value of the ith component in w also represents the contribution of the ith element of each feature vector in the PCA space to classification. Therefore, the analysis of both v and w gives the discriminative region for classification. In Cuingnet et al. (2011), the orthogonal vectors v to the hyperplanes of Thickness-Direct and Thickness-Atlas have been visualized. We also visualized the axis w of LDA by converting it to a pair of vectors on the left and right atlas meshes: w in the PCA space was first transformed to a vector $x = W_P w$ in the feature space. The vector is then divided into two parts: frequency components $\{f_1^L,\ldots,f_F^L\}$ for the left atlas mesh and frequency components $\{f_1^R,\ldots,f_F^R\}$ for the right atlas mesh. These frequency components are finally transformed to two cortical thickness vectors on the left and right atlas meshes using Eq. (3), respectively. We divided these vectors by their magnitudes to obtain two unit vectors for visualization. Fig. 8 depicts these vectors on the atlas meshes for HC vs. AD classification, MCInc vs. MCIc classification, and HC vs. MCIc classification. The entorhinal cortex was the most discriminative for AD classification, and the lateral temporal lobe and the prefrontal cortex were also discriminative, which is consistent with discriminative regions of Thickness-Direct and Thickness-Atlas.

In general, near-by regions in a cortex have correlated brain functions, and are similarly deformed by brain diseases. By representing the noise-filtered cortical thickness data of a subject in terms of the spatial frequency components, our method reflects spatial coherency of the data, which resulted in the spatial coherency of discriminative regions. Our discriminative regions were therefore smoother than those of Thickness-Direct: The vertex-wise cortical thickness representation adopted by Thickness-Direct poorly reflects spatial relationship of the feature data. On the other hand, Thickness-Atlas is based on a region-wise representation of the thickness data, which poorly reflects detailed spatial variation of the thickness data. Due to this property of Thickness-Atlas, the cortical thickness data at all vertices in the same parcellated region contributed equally to the classification (see Fig. 6 in Cuingnet et al. (2011)).

Incremental classification performance

In this section, we demonstrated the effectiveness of incremental classification in both accuracy and efficiency. We initialized three individual subject classifiers, an HC vs. AD classifier, an MCInc vs. MCIc classifier, an HC vs. MCIc classifier using the training data of data

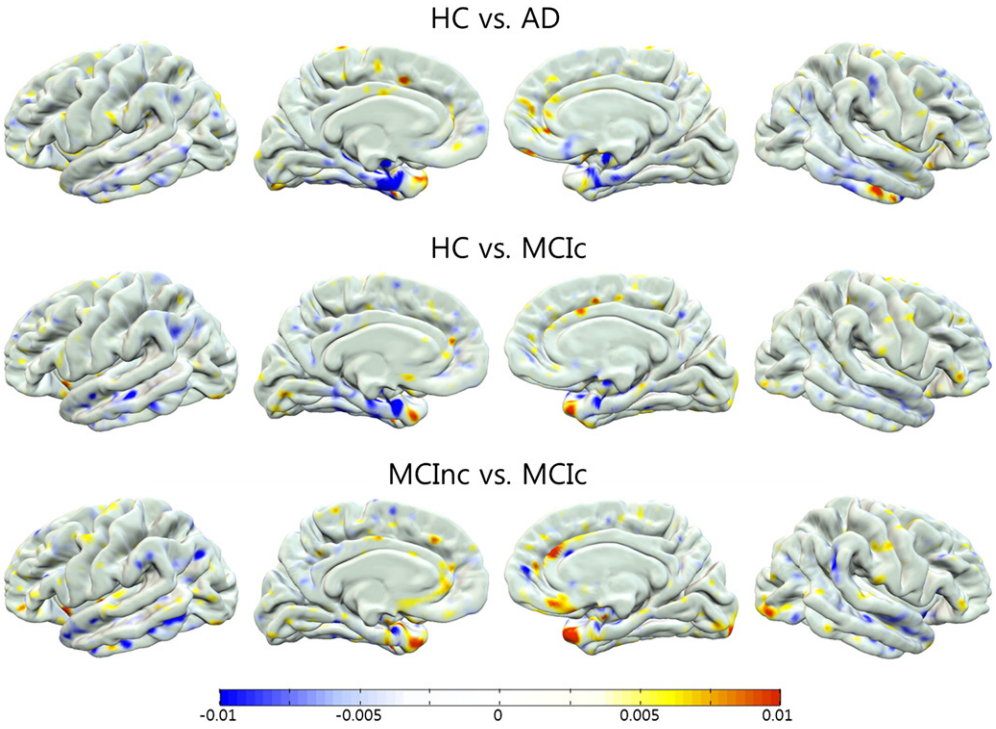

Fig. 8. The discriminative regions in HC vs. AD classification, MCInc vs. MCIc classification, and HC vs. MCIc classification: Each figure visualizes the LDA axes on the atlas meshes.

subsets, Datasets 1, 2, and 3, respectively. The resulting classifiers were validated with the test data in their respective data subsets.

We began with demonstrating the accuracy of incremental classification. As the training data to each of three individual subject classifiers, we generated one hundred random permutations of the entire training subjects in the corresponding data subset in order to keep the experiment unbiased to a specific ordering of the subjects in the data subset. For each permutation, the accuracy of the classifier was measured as follows: we applied the classifier to the test data in the data subset while updating the classifier incrementally with the training data. Specifically, we iteratively supplied the training data of ten subjects at a time for incremental learning until all the training data were used up. Whenever the classifier was updated with these data, the entire test data were used to estimate the accuracy of the classifier. By averaging the results over all permutations, we measured the accuracy of the classifier with respect to the number of used training subjects. As plotted in Fig. 9, the accuracy of every individual subject classifier tended to converge to that of the respective group classifier trained with the entire training data in the corresponding data subset as the number of used training subjects approached to that of the training subjects in the data subset.

We next validated the time efficiency of incremental classification by employing the group and individual subject classifiers for HC vs. AD classification. Similarly results would be obtained for MCInc vs.

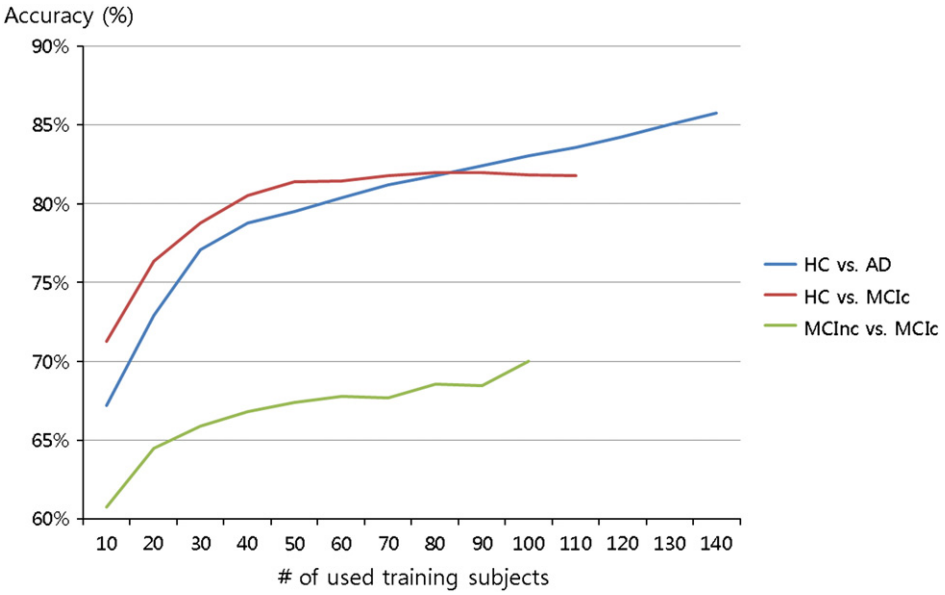

Fig. 9. Average accuracies of individual subject classifiers with respect to the number of used training subjects: The average accuracy of each classifier tended to increase in the number of used training subjects.

MCIc classification and HC vs. MCIc classification. The group classifier was initially trained with the training data of Dataset 1, and the individual subject classifier was initialized with this group classifier. We increased the size of our data set up to (including) 13,000 by cloning the test data of Dataset 1.

We separately measured the computation times for the group and incremental subject classifiers by incrementally supplying test subjects to the both classifiers in the unit of 100 subjects. Both classifiers were updated whenever new training data were added. Fig. 10 shows the computation time of each classifier excluding that for the feature vector construction. The computation time for incremental learning was constantly 1.4 s regardless of the cumulative size of training data since it is dependent on the size of new training data (or new labeled test data). On the other hand, the computation time of batch learning is rapidly growing in the cumulative size of the training data. The experiment of batch learning with more than 13,000 subjects was unable to be performed due to lack of memory space.

Discussion

In this paper, we presented an individual subject classification method based on incremental learning for AD diagnosis and AD prediction using cortical thickness data. We represented cortical thickness data in a frequency domain by employing the MHT. The basis functions for the MHT were from the eigenfunctions of the LB operator which is dependent only on the geometry of a cortical surface but not on the cortical thickness defined on it. Even with vertex-wise features, our method was robust to noise by filtering out high frequency components of cortical thickness data while reflecting their spatial variation. The method not only classified individual subjects with high accuracy, but also enhanced performance of the classifier incrementally. Through experiments, the method demonstrated high performance in both AD diagnosis and AD prediction.

Our classification method provides a general framework to classify individual subjects using arbitrary features defined on a 3D cortical surface. The method can be employed to diagnose and predict other brain diseases using different neuroanatomical or geometrical features such as local surface deformity and curvature. In the future, we would like to apply the framework to subcortical structures for diagnosis of various neurological and psychiatric diseases such as schizophrenia or autism. However, both of PCA and LDA work effectively under the assumption of Gaussian data distributions. Moreover, the mean is assumed to be the discriminating factor rather than the variance in LDA. Therefore, our method is not guaranteed to work effectively for data sets with non-Gaussian distributions or with their variances more discriminative than their means.

Our incremental classification scheme is based on the assumption that the atlas surface (or template surface) is fixed over all subjects. In order to satisfy this assumption, the cortical thickness should be measured using the same template surface, for example, the one available in the FreeSurfer software. A more natural solution is to allow population-specific templates, which we leave as a future research topic. We used the goodness of fit G as the criterion to determine the cut-off dimension F. We conservatively set $G = 0.025$ to obtain $F = 2400$. However, Fig. 4 shows lower cut-off dimensions with better accuracies for different classifiers. Therefore, the accuracy of a classifier could be improved by optimally choosing the goodness of fit in a classifier-specific manner. A similar argument can be applied to dimension reduction of the PCA space. Finally, our classification scheme is based on LDA and PCA for clarity of presentation and easy extension to incremental learning. As mentioned in Group classifier training section, the classification performance could be improved by employing more sophisticated statistical techniques such as PLS and OPLS.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2011-0018262).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the

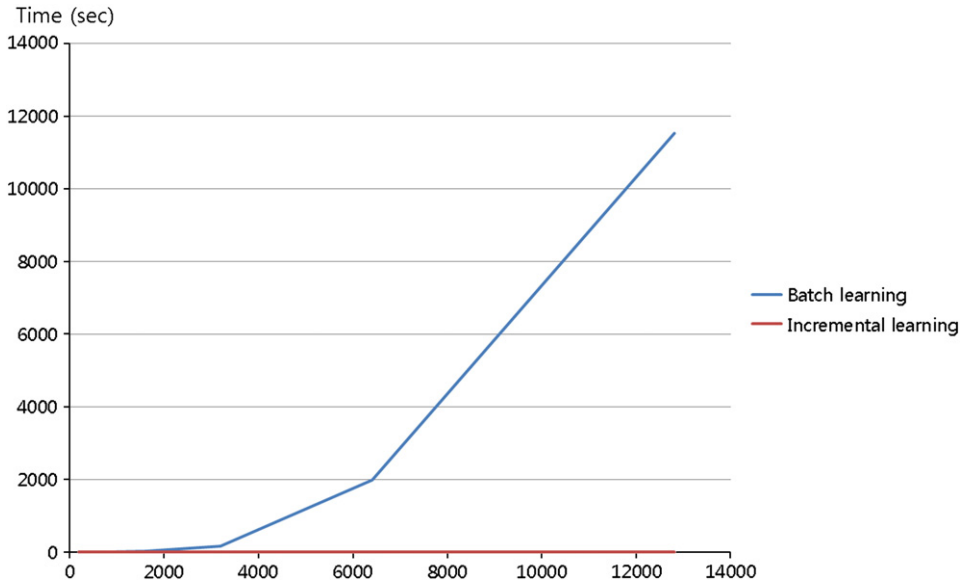

Fig. 10. The computation times for training the HC vs. AD classifier in batch learning and incremental learning: We increased the size of our data up to (including) 13,000 by reusing the test data of Dataset 1. We separately measured the computation times for the incremental learning and the batch learning, excluding that for the feature vector construction. The computation time for incremental learning was constantly 1.4 s regardless of the cumulative size of training data since it is dependent on the size of new training data. On the other hand, the computation time of batch learning rapidly growing in the cumulative size of training data.

following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc., F. Hoffman-La Roche, Schering-Plough, Synarc, Inc., as well as non-profit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to ADNI are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). 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 discriminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129, K01 AG030514, and the Dana Foundation.

Appendix A

We explain how to find an axis maximizing Eq. (7):

$$J(w) = \frac{\sigma_{inter}(w)}{\sigma_{intra}(w)} = \frac{w^T S_B w}{w^T S_W w}.$$

In order to find the axis w that maximizes $J(w)$, we take the derivative with respect to w and set the results to zero:

$$\frac{d}{dw}J(w) = \frac{\left(\frac{d}{dw}w^t S_B w\right) w^t S_W w - \left(\frac{d}{dw}w^t S_W w\right) w^t S_B w}{\left(w^t S_W w\right)^2} = \frac{(2S_B w) w^t S_W w - (2S_W w) w^t S_B w}{\left(w^t S_W w\right)^2} = 0.$$

This equation can be simplified as follows:

$$\begin{aligned} & w^t S_W w (S_B w) - w^t S_B w (S_W w) = 0 \\ & \frac{w^t S_W w (S_B w)}{w^t S_W w} - \frac{w^t S_B w (S_W w)}{w^t S_W w} = 0 \\ & S_B w - \frac{w^t S_B w}{w^t S_W w} S_W w = 0 \\ & S_B w = \lambda S_W w \\ & S_W^{-1} S_B w = \lambda w. \end{aligned}$$

Since $\lambda = \frac{w^t S_B w}{w^t S_W w}$ is the energy function $J(w)$, the axis maximizing $J(w)$ is the eigenvector of $S_W^{-1} S_B$ with the largest eigenvalue.

References

Anticevic, A., Dierker, D.L., Gillespie, S.K., Repovs, G., Csernansky, J.G., Essen, D.C.V., Barch, D.M., 2008. Comparing surface-based and volume-based analyses of functional neuroimaging data in patients with schizophrenia. Neuroimage 41 (3), 835–848.

Bain, L.J., Jedrziewski, K., Morrison-Bogorad, M., Albert, M., Cotman, C., Hendrie, H., Trojanowski, J.Q., 2008. Healthy brain aging: a meeting report from the Sylvan M. Cohen annual retreat of the University of Pennsylvania Institute on Aging. Alzheimers Dement. 4 (6), 443–446.

Balakrishnama, S., Ganapathiraju, A., 1998. Linear discriminant analysis — a brief tutorial [online] (Available:) http://www.music.mcgill.ca/ich/classes/mumt611_05/classifiers/lda_theory.pdf1998.

Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J., 1997. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans. Pattern Anal. Mach. Intell. 19, 711–720.

Bylesjø, M., Rantalainen, M., Cloarec, O., Nicholson, J.K., Holmes, E., Tryg, J., 2006. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. J. Chemom. 20 (8–10), 341–351.

Chételat, G., Landeau, B., Eustache, F., Mézenge, F., Viader, F., de la Sayette, V., Desgranges, B., Baron, J.-C., 2005. Using voxel-based morphometry to map the structural changes associated with rapid conversion in MCI: a longitudinal MRI study. Neuroimage 27 (4), 934–946.

Chung, M., Dalton, K., Li, S., Evans, A., Davidson, R., 2007. Weighted Fourier series representation and its application to quantifying the amount of gray matter. IEEE Trans. Med. Imaging 26 (4), 566–581 (April).

Chupin, M., Mukuna-Bantumbakulu, A.R., 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.

Colliot, O., Chételat, G., Chupin, M., Desgranges, B., Magnin, B., Benali, H., Dubois, B., Garnero, L., Eustache, F., Lehéricy, S., 2008. Discrimination between Alzheimer Disease, mild cognitive impairment, and normal aging by using automated segmentation of the hippocampus. Radiology 248 (1), 194–201.

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. Corrected Proof. 56 (2), 766–781.

Desai, R., Liebenthal, E., Possing, E.T., Waldron, E., Binder, J.R., 2005. Volumetric vs. surface-based alignment for localization of auditory cortex activation. Neuroimage 26 (4), 1019–1029.

Desikan, R.S., Cabral, H.J., Hess, C.P., Dillon, W.P., Glastonbury, C.M., Weiner, M.W., Schmansky, N.J., Greve, D.N., Salat, D.H., Buckner, R.L., Fischl, B., Initiative, A.D.N., 2009. Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. Brain 132 (8), 2048–2057.

Dickerson, B.C., Bakkour, A., Salat, D.H., Feczko, E., Pacheco, J., Greve, D.N., Grodstein, F., Wright, C.I., Blacker, D., Rosas, H.D., Sperling, R.A., Atri, A., Growdon, J.H., Hyman, B.T., Morris, J.C., Fischl, B., Buckner, R.L., 2009. The cortical signature of Alzheimer's disease: regionally specific cortical thinning relates to symptom severity in very mild to mild AD dementia and is detectable in asymptomatic amyloid-positive individuals. Cereb. Cortex 19 (3), 497–510.

Dubois, B., Albert, M.L., 2004. Amnestic MCI or prodromal Alzheimer's disease? Lancet Neurol. 3 (4), 246–248.

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

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. 7 (7), 179–188.

Folstein, M.F., Folstein, S.E., McHugh, P.R., 1975. Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. J. Psychiatr. Res. 12 (3), 189–198.

Gerardin, E., Châtelat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.-S., Niethammer, M., Dubois, B., Lehéricy, S., Garnero, L., Eustache, F., Colliot, O., 2009. Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. Neuroimage 47 (4), 1476–1486.

Good, C.D., Scahill, R.I., Fox, N.C., Ashburner, J., Friston, K.J., Chan, D., Crum, W.R., Rossor, M.N., Frackowiak, R.S., 2002. Automatic differentiation of anatomical patterns in the human brain: validation with studies of degenerative dementias. Neuroimage 17 (1), 29–46.

Hall, P.M., Marshall, D., Martin, R.R., 1998. Incremental eigenanalysis for classification. British Machine Vision Conference, pp. 286–295.

Hall, P., Marshall, D., Martin, R., 2002. Adding and subtracting eigenspaces with eigenvalue decomposition and singular value decomposition. Image Vision Comput. 20 (13–14), 1009–1016.

Jack Jr., C., Knopman, D.S., Jagust, W.J., Shaw, L.M., Aisen, P.S., Weiner, M.W., Petersen, R.C., Trojanowski, J.Q., 2010. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol. 9 (1), 119–128.

Jolliffe, I.T., 2002. Principal Component Analysis, 2nd ed. Springer. (Oct).

Karas, G.B., Burton, E.J., Rombouts, S.A.R.B., van Schijndel, R.A., O'Brien, J.T.h, Scheltens, P., McKeith, I.G., Williams, D., Ballard, C., Barkhof, F., 2003. A comprehensive study of gray matter loss in patients with Alzheimer's disease using optimized voxel-based morphometry. Neuroimage 18 (4), 895–907.

Khan, A.R., Wang, L., Beg, M.F., 2008. Freesurfer-initiated fully-automated subcortical brain segmentation in mri using large deformation diffeomorphic metric mapping. Neuroimage 41 (3), 735–746.

Kim, S.-G., Chung, M., Seo, S., Schaefer, S., Reekum, C., Davidson, R., accepted for publication. Heat kernel smoothing via Laplace–Beltrami eigenfunctions and its application to subcortical structure modeling, in: Pacific-Rim Symposium on Image and Video Technology (PSIVT). Lecture Notes in Computer Science (LNCS).

Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack, C.R., Ashburner, J., Frackowiak, R.S.J., 2008. Automatic classification of MR scans in Alzheimer's disease. Brain 131 (3), 681–689.

Lerch, J.P., Pruessner, J.C., Zijdenbos, A., Hampel, H., Teipel, S.J., Evans, A.C., 2005. Focal decline of cortical thickness in Alzheimer's disease identified by computational neuroanatomy. Cereb. Cortex 15 (7), 995–1001 (July).

Levy, B., 2006. Laplace–Beltrami eigenfunctions towards an algorithm that "understands" geometry. SMI'06: Proceedings of the IEEE International Conference on Shape Modeling and Applications 2006. IEEE Computer Society, Washington, DC, USA, p. 13.

Levy, A., Lindenbaum, M., 2000. Sequential Karhunen–Loeve basis extraction and its application to images. IEEE Trans. Image Process. 9, 1371–1374.

Lim, J., Ross, D., sung Lin, R., hsuan Yang, M., 2004. Incremental learning for visual tracking. Advances in Neural Information Processing Systems. MIT Press, pp. 793–800.

Liu, Y., Rayens, W., 2007. Pls and dimension reduction for classification (Jul) Comput. Stat. 22 (2), 189–208 (URL) http://dx.doi.org/10.1007/s00180-007-0039-y.

Liu, C., Wechsler, H., 2000. Robust coding schemes for indexing and retrieval from large face databases. IEEE Trans. Image Process. 9 (1), 132–137.

Magnin, B., Mesrob, L., Kinkingnehun, S., Issac, P.-M., Colliot, O., Sarazin, M., Dubois, B., Lehericy, S., Benali, H., 2009. Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. Neuroradiology 51, 73–83.

Misra, C., Fan, Y., Davatzikos, C., 2009. Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to ad: results from ADNI. Neuroimage 44 (4), 1415–1422.

Pang, S., Ozawa, S., Kasabov, N., 2005. Incremental linear discriminant analysis for classification of data streams. IEEE Trans. Syst. Man Cybern. B Cybern. 35 (5), 905–914 (Oct).

Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Tangalos, E.G., Kokmen, E., 1999. Mild cognitive impairment: clinical characterization and outcome. Arch. Neurol. 56 (3), 303–308.

Qiu, A., Miller, M.I., 2008. Multi-structure network shape analysis via normal surface momentum maps. Neuroimage 42 (4), 1430–1438.

Qiu, A., Bitouk, D., Miller, M., 2006. Smooth functional and structural maps on the neocortex via orthonormal bases of the Laplace–Beltrami operator. IEEE Trans. Med. Imaging 25 (10), 1296–1306 (oct).

Qiu, A., Younes, L., Miller, M.I., Csernansky, J.G., 2008. Parallel transport in diffeomorphisms distinguishes the time-dependent pattern of hippocampal surface deformation due to healthy aging and the dementia of the Alzheimer's type. Neuroimage 40 (1), 68–76.

Querbes, O., Aubry, F., Pariente, J., Lotterie, J.-A., Demonet, J.-F., Duret, V., Puel, M., Berry, I., Fort, J.-C., Celsis, P., Initiative, T.A.D.N., 2009. Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. Brain 132 (8), 2036–2047.

Ross, D.A., Lim, J., Lin, R.-S., Yang, M.-H., 2008. Incremental learning for robust visual tracking. Int. J. Comput. Vis. 77, 125–141 (May).

Seo, S., Chung, M., 2011. Laplace–Beltrami eigenfunction expansion of cortical manifolds. IEEE International Symposium on Biomedical Imaging.

Seo, S., Chung, M., Voperian, H., 2010. Heat kernel smoothing using Laplace–Beltrami eigenfunctions. 13th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI): Lecture Notes in Computer Science (LNCS), pp. 505–512.

Seo, S., Chung, M., Voperian, H., 2011. Mandible shape modeling using the second eigenfunction of the Laplace–Beltrami operator. SPIE Med. Imaging 7962, 79620Z–79620Z-6.

Shen, L., Huang, H., Makedon, F., Saykin, A.J., 2007. Efficient registration of 3D SPHARM Surfaces. CRV'07: Proceedings of the Fourth Canadian Conference on Computer and Robot Vision, pp. 81–88.

Singh, V., Mukherjee, L., Chung, M.K., 2008. Cortical surface thickness as a classifier: boosting for autism classification. Proceedings of the 11th International Conference on Medical Image Computing and Computer-Assisted Intervention — Part I. MICCAI'08, pp. 999–1007.

Thompson, P.M., Hayashi, K.M., de Zubicaray, G., Janke, A.L., Rose, S.E., Semple, J., Herman, D., Hong, M.S., Dittmer, S.S., Doddrell, D.M., Toga, A.W., 2003. Dynamics of gray matter loss in Alzheimer's disease. J. Neurosci. 23 (3), 994–1005.

Vallet, B., Lévy, B., 2008. Spectral geometry processing with manifold harmonics. Computer Graphics Forum (Proceedings Eurographics).

Wang, L., Miller, J.P., Gado, M.H., McKeel, D.W., Rothermich, M., Miller, M.I., Morris, J.C., Csernansky, J.G., 2006. Abnormalities of hippocampal surface structure in very mild dementia of the Alzheimer type. Neuroimage 30 (1), 52–60.

Ye, J., Chen, K., Wu, T., Li, J., Zhao, Z., Patel, R., Bae, M., Janardan, R., Liu, H., Alexander, G., Reiman, E., 2008. Heterogeneous data fusion for Alzheimer's disease study. KDD'08: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 1025–1033.

Yu, H., Yang, J., 2001. A direct LDA algorithm for high-dimensional data with application to face recognition. Pattern Recognit. 34, 2067–2070.

Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A., 2003. Face recognition: a literature survey. ACM Comput. Surv. 35, 399–458 (December).

Zhu, M., 2006. A study of the generalized eigenvalue decomposition in discriminant analysis. Ph.D. thesis, The Ohio State University.