NeuroImage 92 (2014) 322–339

Contents lists available at ScienceDirect

NeuroImage

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

Brain extraction based on locally linear representation-based classification

Meiyan Huang [1], Wei Yang [1], Jun Jiang, Yao Wu, Yu Zhang, Wufan Chen, Qianjin Feng *, for the Alzheimer's Disease Neuroimaging Initiative [2]

School of Biomedical Engineering, Southern Medical University, Guangzhou, China

ARTICLE INFO

Article history:
Accepted 31 January 2014
Available online 10 February 2014

Keywords:
Brain extraction
Locally Linear Representation-based Classification
Label fusion
Local anchor embedding

ABSTRACT

Brain extraction is an important procedure in brain image analysis. Although numerous brain extraction methods have been presented, enhancing brain extraction methods remains challenging because brain MRI images exhibit complex characteristics, such as anatomical variability and intensity differences across different sequences and scanners. To address this problem, we present a Locally Linear Representation-based Classification (LLRC) method for brain extraction. A novel classification framework is derived by introducing the locally linear representation to the classical classification model. Under this classification framework, a common label fusion approach can be considered as a special case and thoroughly interpreted. Locality is important to calculate fusion weights for LLRC; this factor is also considered to determine that Local Anchor Embedding is more applicable in solving locally linear coefficients compared with other linear representation approaches. Moreover, LLRC supplies a way to learn the optimal classification scores of the training samples in the dictionary to obtain accurate classification. The International Consortium for Brain Mapping and the Alzheimer's Disease Neuroimaging Initiative databases were used to build a training dataset containing 70 scans. To evaluate the proposed method, we used four publicly available datasets (IBSR1, IBSR2, LPBA40, and ADNI3T, with a total of 241 scans). Experimental results demonstrate that the proposed method outperforms the four common brain extraction methods (BET, BSE, GCUT, and ROBEX), and is comparable to the performance of BEaST, while being more accurate on some datasets compared with BEaST.

Introduction

Brain extraction, also known as skull stripping, aims to remove non-brain tissues (e.g., scalp, skull, and dura); this procedure is an important step in brain image analysis. Stripped MRI brain images provide several advantages in terms of several factors, such as brain tissue classification (Shattuck et al., 2001), registration (Shen and Davatzikos, 2004), and cortical surface reconstruction (Dale et al., 1999). Accurate brain extraction is also important for cortical thickness estimation; on the one hand, cortical thickness may be overestimated if the dura is not removed (van der Kouwe et al., 2008). On the other hand, cortical thickness may be underestimated if the cortical surface is unintentionally removed. The manual delineation of the brain is time consuming and suffers from inter-operator variations. For these reasons, semi-automated and automated brain extraction methods are more preferred than manual delineation.

Anatomical changes in the brain caused by diseases or old age present a major challenge when designing a brain extraction method. For instance, the brains of older individuals usually exhibit atrophy with higher rates of brain tissue loss compared with those of younger individuals. Diseases such as Alzheimer's disease (AD) and mild cognitive impairment (MCI) lead to the loss of brain tissue. Image variations also present another challenge because of various acquisition sequences and scanner types. Most existing brain extraction methods often need to be tuned to work on a certain type of study or a certain population. Hence, a reliable and robust method that is capable of working on a variety of brain morphologies and acquisition sequences would be highly desired in neuroimaging studies.

To extract the brain, researchers developed numerous algorithms, such as morphology operations (Chiverton et al., 2007; Lemieux et al., 1999; Mikheev et al., 2008; Park and Lee, 2009; Ward, 1999), atlas matching (Ashburner and Friston, 2000), histogram analysis

* Corresponding author at: School of Biomedical Engineering, Southern Medical University, Guangzhou 510515, China. Fax: +86 20 61648286.

E-mail addresses: huangmeiyan11@gmail.com (M. Huang), weiyanggm@gmail.com (W. Yang), smujiang@gmail.com (J. Jiang), wuyao198851@gmail.com (Y. Wu), gz.akita.zy@gmail.com (Y. Zhang), wufanchen@gmail.com (W. Chen), qianjinfeng08@gmail.com (Q. Feng).

[1] These authors contributed equally to this work.

[2] 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. Complete listing of ADNI investigators is available at http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Authorship_List.pdf.

1053-8119/$ – see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2014.01.059

(Shan et al., 2002), watershed (Hahn and Peitgen, 2000), graph cuts (Sadananthan et al., 2010), level sets (Baillard et al., 2001; Zhuang et al., 2006), deformable models (Smith, 2002), label fusion (Eskildsen et al., 2012; Leung et al., 2011), and hybrid approaches (Carass et al., 2011; Iglesias et al., 2011; Rehm et al., 2004; Rex et al., 2004; Segonne et al., 2004; Shattuck et al., 2001; Shi et al., 2012). Each of these methods provides advantages and disadvantages. For instance, morphology operations are fast and can be easily adjusted; however, this method fails to determine the optimum morphology size necessary to separate brain tissues from non-brain tissues (Park and Lee, 2009). Histogram (Shan et al., 2002) and watershed (Hahn and Peitgen, 2000) methods are simple and consistently producing complete boundaries. However, these two methods are sensitive to noise, which is a common problem encountered in intensity-based methods. Brain extraction methods based on deformable surfaces can achieve a smooth closed surface. However, these methods assume that the brain surface is smooth with low curvature; this characteristic is often not observed on the brain boundary, particularly in basal regions (Hahn and Peitgen, 2000). In meta-algorithm (Rex et al., 2004; Shi et al., 2012), several existing brain extraction methods are combined to compensate for the weaknesses of each method. However, the model should be specifically designed through meta-algorithm to gain optimum performance when new data are sufficiently different from previous training datasets (Rex et al., 2004).

Label fusion-based segmentation methods have been extensively studied. For instance, in MAPS (Leung et al., 2011), non-rigid registrations of selected atlases to the target image are initially used, and a label fusion technique is then applied to merge the labels from the atlases to create an optimal segmentation in the target image. SuperDyn (Khan et al., 2011) is another popular fusion-based method, in which the spatially local weights for atlases are determined by combining the supervised weight learned from the training set and the dynamic weight obtained from the target–atlas pairing. The main disadvantages of these two methods include (1) their long computational time (19 h for MAPS), and (2) the heavy dependence of the segmentation performance on the registration accuracy. A patch-based label fusion method called BEaST has been proposed (Eskildsen et al., 2012). In BEaST, non-rigid registration is not required but is replaced with rough affine alignment to reduce computational costs. The weights of the fused labels are calculated using non-local means approach (Buades et al., 2005); experimental results show that the patch-based label fusion approach significantly increases the segmentation accuracy. Furthermore, a multi-resolution framework is used in BEaST to improve computational efficiency and robustness. Although label fusion-based methods provide optimum performance for brain extraction, a number of fundamental problems of label fusion, such as the estimation of the labels of test samples by linearly combining the labels of training samples and the mechanism by which fusion weights are calculated remain unclear and require further investigation.

In the current study, a Locally Linear Representation-based Classification (LLRC) method for brain extraction is presented. In LLRC, the locally linear representation is introduced into the classical classification model and a novel classification framework is derived. Under this classification framework, the label fusion approach can be considered a special case and thoroughly interpreted. Locality is important to calculate fusion weights for LLRC; this factor is also considered to determine that Local Anchor Embedding (LAE) (Liu et al., 2010) is more applicable in solving locally linear coefficients compared with other linear representation approaches, such as Sparse Coding (SC) (Wright et al., 2009), non-local means (Buades et al., 2005), and Locality-constrained Linear Coding (LLC) (Wang et al., 2010). Moreover, LLRC supplies a way to learn the optimal classification scores of the training samples in the dictionary to obtain accurate classification. The proposed method was tested on multiple datasets acquired on different scanners. The performance of the proposed brain extraction method was thoroughly evaluated by comparing with other methods, such as brain extraction tool (BET) (Smith, 2002), brain surface extractor (BSE) (Shattuck et al., 2001), GCUT (Sadananthan et al., 2010), ROBEX (Iglesias et al., 2011), and BEaST.

Datasets

Six public datasets (two for training and four for evaluation) were used in our study. The scan parameters of each dataset are listed in Table 1.

The training dataset consisted of 70 T1-weighted scans from two datasets, in which 10 scans were obtained from the International Consortium for Brain Mapping (ICBM) database (age: 18 years to 43 years) (Mazziotta et al., 1995) and 60 scans were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (age: 55 years to 91 years) (Mueller et al., 2005). The ICBM database consisted of healthy subjects. The ADNI database contained cognitive-normal (CN) subjects and subjects with AD and MCI. 20 T1 MRI scans from each class (CN, AD, MCI) were chosen to construct the ADNI training dataset in the present study. All of the scans and their corresponding brain masks in the training dataset were obtained from the websites found in a previous study (Eskildsen et al., 2012). To increase the size of our training dataset, we flipped these 70 scans and their corresponding brain masks along the mid-sagittal plane by utilizing the symmetric properties of the human brain. Thus, our training dataset consisted of a total of 140 scans (original and flipped).

The first test dataset, called IBSR1, was provided by the Internet Brain Segmentation Repository (IBSR)[3] and consisted of 18 T1-weighted scans and their corresponding brain masks obtained from healthy subjects (age: 7 years to 71 years). Some of the scans showed relatively low contrast between the brain and surrounding tissues.

The second test dataset, also provided by IBSR, was named IBSR2 and comprised 20 T1-weighted scans of normal subjects (29.0 ± 4.8 years old) and their corresponding brain masks. This dataset exhibited low resolution in addition to high heterogeneity of several scans; as a result, classifying IBSR2 was challenging.

The third test dataset, namely, LPBA40[4], consisted of 40 T1-weighted scans of normal subjects (29.2 ± 6.30 years old) and their corresponding brain masks.

The fourth test dataset (ADNI3T dataset) consisted of 163 (46 CN, 80 MCI, and 37 AD) T1-weighted MRI scans and their corresponding brain masks from the baseline time point of the ADNI database. The demographics of the subjects are shown in Table 2.

The manual brain extraction protocols of these datasets are given as follows:

For the training dataset, the brain mask includes all cerebral and cerebellar white matter (WM), all cerebral and cerebellar gray matter (GM), cerebral spinal fluid (CSF) in the ventricles (lateral, third and fourth) and cerebellar cistern, CSF in deep sulci and along the surface of the brain and brain stem, and the brainstem (pons, medulla).

For the IBSR1 and IBSR2 test datasets, the brain mask includes all cerebral and cerebellar WM, all cerebral and cerebellar GM, CSF in the ventricles (lateral, third and fourth), CSF in deep sulci and along the surface of the brain and brain stem, and the brainstem (pons, medulla).

The definition of the brain mask of the LPBA40 test dataset is the same as that of the training dataset.

For the ADNI3T test dataset, the brain mask includes GM and WM and excludes internal and external CSF.

[3] http://www.cma.mgh.harvard.edu/ibsr/

[4] http://www.loni.ucla.edu/

Table 1
Scan parameters of each dataset.

Dataset		Field strength (T)	Imaging parameters
Training dataset	ICBM	1.5	TR = 18 ms, TE = 10 ms, flip angle = 30°, and rectangular fields of view of 256 mm (superior–inferior) and 204 mm (anterior–posterior). The sagittal slice thickness was 1 mm.
	ADNI	1.5	TR = 2300 ms, TI = 1000 ms, TE = 3.5 ms, and flip angle = 8°. The sagittal slice thickness was 1.2 mm with an in-plane resolution of 0.94 mm.
Test dataset	IBSR1	1.5	Each scan has 128 coronal slices with pixel dimensions ranging from 0.84 mm to 1 mm on each slice. The slice thickness is 1.5 mm.
	IBSR2	1.5	Ten scans: TR = 40 ms, TE = 8 ms, and flip angle = 50°. The slice thickness of each scan was 3.1 mm and the in-plane resolution is 1 mm × 1 mm.
			Ten scans: TR = 50 ms, TE = 9 ms, and flip angle = 50°. The slice thickness of each scan was 3.0 mm and the in-plane resolution is 1 mm × 1 mm.
	LPBA40	1.5	TR = 10.0 ms to 12.5 ms; TE = 4.22 ms to 4.5 ms; and flip angle = 20°. The coronal slice thickness was 1.5 mm with an in-plane resolution of 0.86 mm (38 subjects) or 0.78 mm (2 subjects).
	ADNI3T	3	TR = 2300 ms, TI = 900 ms, minimum full TE, and flip angle = 8°. The sagittal slice thickness was 1.2 mm with an in-plane resolution of 1 mm.

Methods

The proposed method consists of four major steps: preprocessing, feature extraction, LLRC-based classification, and postprocessing. To reduce the computational costs, we embedded the proposed method in a multi-resolution framework. The flowchart of the proposed method is illustrated in Fig. 1.

LLRC

Basic idea of LLRC

Brain extraction can be considered as a two-class (brain tissue and non-brain tissue) classification problem. A two-class classification problem is defined in this section.

For a test sample $\boldsymbol{x} \in R^K$, the objective of this classification is to find the function $H : R^K \mapsto R$ such that $H(\boldsymbol{x})$ is a "good" predictor of a real classification score $y \in [0, 1]$, where the sample $\boldsymbol{x}$ stands for the image feature in the current study (details in the "Feature extraction" section) and the classification score y stands for the probability that sample $\boldsymbol{x}$ belongs to the object. To learn H, we used a training set $\mathbf{T} = \{\boldsymbol{x}^{(i)}, y^{(i)}\}_{i=1}^{M}$, where M known sample-score pairs are included. The cost function of this classification can be defined based on this training set by using

$$J = \sum\nolimits_{i=1}^{M} \left\| H\left(\boldsymbol{x}^{(i)}\right) - y^{(i)} \right\|^2. \tag{1}$$

This is a classic learning based classification model. In the current study, locally linear representation was introduced to this model and the LLRC method was proposed. Before the proposed LLRC was introduced, the following assumptions were considered as the bases for LLRC:

Assumption 1. Samples lie on a non-linear manifold that can be approximated locally and linearly on the basis of their nearest neighbors. This assumption is reasonable and has been applied successfully in previous methods (Roweis and Saul, 2000; Wang et al., 2010).

Assumption 2. The classification function H is differentiable and can be fitted locally and linearly.

Table 2
The demographics of the 163 subjects with 3 T MRI scans.

	CN (n = 46)	MCI (n = 80)	AD (n = 37)
Mean age ± SD, years	75.0 ± 4.0	75.3 ± 7.6	74.1 ± 8.9
Gender (male, %)	18 (39%)	48 (60%)	14 (37.8%)

Considering Assumption 1, we can represent sample $\boldsymbol{x}$ as Eq. (2) if a matrix $\boldsymbol{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_N]$, called dictionary in this study, consists of N typical samples of the original sample space:

$$\begin{gathered} \boldsymbol{x} = \boldsymbol{Da} + \boldsymbol{\varepsilon} = \sum\nolimits_{j=1}^{N} a_j \boldsymbol{d}_j + \boldsymbol{\varepsilon} \\ s.t. \|\boldsymbol{\varepsilon}\| < \tau, \\ \forall \boldsymbol{d}_j \notin \mathrm{N}_{\boldsymbol{x}}(k), a_j = 0 \end{gathered} \tag{2}$$

where $\mathrm{N}_{\boldsymbol{x}}(k)$ is the set of k nearest neighbors of $\boldsymbol{x}$ in $\boldsymbol{D}$. $\boldsymbol{\varepsilon}$ is the reconstruction error. τ is a small positive real number that ensures the reconstruction accuracy. $\boldsymbol{a} = (a_1, a_2, \cdots, a_N)^T$ is the weight coefficient vector of the linear combination. Many (at least N-k) elements of $\boldsymbol{a}$ are zeros for an arbitrary sample $\boldsymbol{x}$, indicating that the described linear representation is sparse. A stronger locality constraint should be used in Eq. (2) compared with the previous constraint, such that Eq. (3) is expressed to emphasize the locality of the proposed linear representation and apply Assumption 2 smoothly.

$$\begin{gathered} \boldsymbol{x} = \boldsymbol{Da} + \boldsymbol{\varepsilon} = \sum\nolimits_{j=1}^{N} a_j \boldsymbol{d}_j + \boldsymbol{\varepsilon} \\ s.t. \|\boldsymbol{\varepsilon}\| < \tau \\ \forall \boldsymbol{d}_j \notin N_{\boldsymbol{x}}(k), a_j = 0 \\ \sum\nolimits_{j}^{N} a_j = 1, a_j \geq 0 \end{gathered} \tag{3}$$

These constraints ensure that the reconstructed sample is a convex combination of its closest neighbors. In other words, the reconstructed sample is located in a small convex region on a hyperplane spanned by the closest neighbors. Considering that this small convex region is consistent with Assumption 2, such that the classification function H is linear in this small local region, we can rewrite $H(\boldsymbol{x})$ (ignoring $\boldsymbol{\varepsilon}$) as Eq. (4):

$$H(\boldsymbol{x}) = H(\boldsymbol{Da}) = H\left(\sum_{j=1}^{N} a_j \boldsymbol{d}_j\right) = \sum_{j=1}^{N} a_j H\left(\boldsymbol{d}_j\right) = \sum_{j=1}^{N} a_j h_j = \boldsymbol{a}^T \boldsymbol{h}, \tag{4}$$

where $\boldsymbol{h} = (h_1, h_2, \cdots, h_N)^T$, and $h_j = H(\boldsymbol{d}_j)$ is the classification score of the jth atom vector in $\boldsymbol{D}$ and should belong to [0,1]. Eq. (4) indicates that the classification score of a sample can be estimated by the linear combination of the classification scores of its k nearest neighbors in $\boldsymbol{D}$. The combination coefficients can be calculated by solving linear representation problem in original sample space ("Locally linear representation" section).

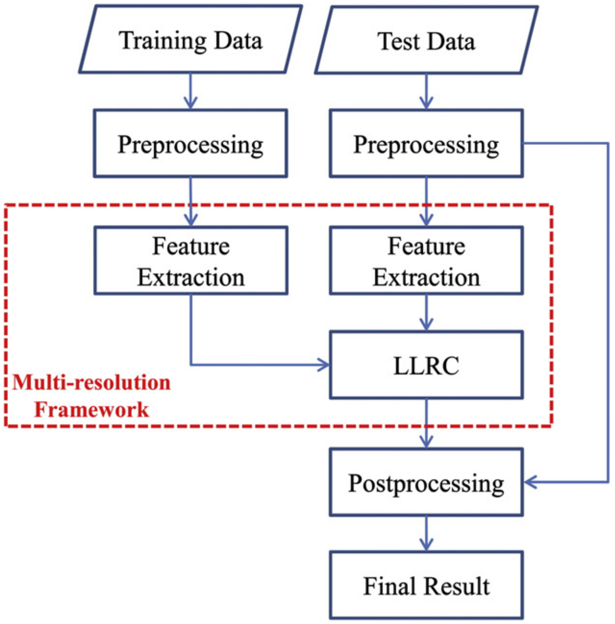

Fig. 1. Flowchart of the proposed method.

On the basis of Eq. (4), the cost function defined by Eq. (1) can be rewritten as

$$J(\boldsymbol{h}) = \sum_{i=1}^{M} \left\| \sum_{j=1}^{N} a_j^{(i)} h_j - y^{(i)} \right\|^2 = \sum_{i=1}^{M} \left\| \boldsymbol{a}^{(i)^T} \boldsymbol{h} - y^{(i)} \right\|^2 s.t. \quad h_j \subset [0, 1], \tag{5}$$

where h_j is considered as a parameter that should be learnt. To solve Eq. (5), we can use the trust-region method ("Classification score learning" section).

The proposed LLRC method was based on these concepts. Using the training set $\mathbf{T} = \{\boldsymbol{x}^{(i)}, y^{(i)}\}_{i=1}^{M}$, we should construct a dictionary $\boldsymbol{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_N]$; we can also estimate $\boldsymbol{h} = (h_1, h_2, \cdots, h_N)^T$, which consists of the classification score of each atom vector in $\boldsymbol{D}$, by minimizing Eq. (5). For a test sample $\boldsymbol{x}$, the classification score of $\boldsymbol{x}$ can be estimated according to Eq. (4) if the corresponding weight coefficient vector $\boldsymbol{a}$ in Eq. (3) can be solved appropriately.

Eq. (4) indicates the basic principle of the label fusion approaches, in which we use the linear combination of the labels (classification score) of the training samples to estimate the label of the test sample. A label fusion approach can be considered a special case of LLRC and therefore is reasonable only when Assumptions 1 and 2 are held. In addition, Assumption 2 suggests that linear representation should be constrained to a local region, indicating that locality is an important factor in solving linear representation.

Dictionary leaning

Manually labeled original samples in a training set are used to construct $\boldsymbol{D}$ (Wright et al., 2009). However, numerous original training samples possibly produce a large $\boldsymbol{D}$, which dramatically increases computational and memory costs. In the current study, more than one million labeled samples are available for training. As such, subsequent processes are impractical when conducted traditionally. In addition, several previous studies (Gao et al., 2012; Wang et al., 2010) have verified that k-means (Macqueen, 1967) produces a representative dictionary for sparse representation. Therefore, we used k-means method to cluster the training samples and selected the cluster centers as the atom vectors to construct $\boldsymbol{D}$.

To obtain $\boldsymbol{D}$, we partitioned the original sample set $\{\boldsymbol{x}^{(i)}\}_{i=1}^{M}$ into N sub-sets $\boldsymbol{G} = \{G_i\}_{i=1}^{N}$ by using k-means, such that the within-cluster sum of squares is minimized:

$$\arg\min_{\boldsymbol{G}} \sum_{i=1}^{N} \sum_{\boldsymbol{x}^{(j)} \in G_i} \left\| \boldsymbol{x}^{(j)} - \boldsymbol{d}_i \right\|^2, \tag{6}$$

where $\boldsymbol{d}_i$ is the mean (cluster center) of the samples in G_i.

Locally linear representation

To approximately represent a sample linearly based on the training samples, researchers proposed several methods, such as SC (Wright et al., 2009), non-local means (Buades et al., 2005), LAE (Liu et al., 2010), and LLC (Wang et al., 2010). Sparse representation emphasizes the sparsity of the representation, in which the lowest number of training samples is used to reconstruct a test sample with minimum reconstruction error. Non-local means, LLC, and LAE focus on locality rather than sparsity by limiting linear codes in a local region. In non-local means method, the weight coefficients are determined according to the distance between the samples by using Gaussian kernels; however, using this method cannot ensure that a minimum reconstruction error defined in Eq. (2) is obtained. In LAE, the weight coefficients are maintained as non-negative values and the sum is equal to one; in this method, the reconstructed sample is a convex combination of its closest neighbors. As mentioned in the "Basic idea of LLRC" section, the locality of the representation is preferred for the proposed LLRC. Therefore, LAE was chosen to solve the linear representation in the current study. To ensure completeness, we briefly describe LAE in the succeeding section.

Considering the concrete task in the present paper, we can rewrite the cost function of LAE as

$$\boldsymbol{a}^* = \arg\min_{\boldsymbol{a}} \left\| \boldsymbol{x} - \sum_{j=1}^{N} a_j \boldsymbol{d}_j \right\|^2 s.t. \quad \forall \boldsymbol{d}_j \notin N_{\boldsymbol{x}}(k), a_j = 0 \sum_{j}^{N} a_j = 1, a_j \geq 0 \tag{7}$$

Three steps were performed to obtain the solution of LAE. (I) k nearest neighbors of test sample $\boldsymbol{x}$ was selected from $\boldsymbol{D}$ and $N_{\boldsymbol{x}}(k)$ was constructed; (II) For $\boldsymbol{d}_j$s that do not belong to $N_{\boldsymbol{x}}(k)$, the associated a_js were set to 0; (III) For remaining $\boldsymbol{d}_j$s that belong to $N_{\boldsymbol{x}}(k)$, their corresponding a_js were calculated using the projected gradient method (Liu et al., 2010).

Classification score learning

In the proposed LLRC, the classification score of each atom vector of $\boldsymbol{D}$ was learnt. The cost function defined in Eq. (5) can be rewritten as

$$\begin{aligned} J(\boldsymbol{h}) &= \sum_{i=1}^{M} \left\| \boldsymbol{a}^{(i)^T} \boldsymbol{h} - y^{(i)} \right\|^2 \\ &= \sum_{i=1}^{M} \left[\left(\boldsymbol{a}^{(i)^T} \boldsymbol{h} \right)^2 - 2\boldsymbol{a}^{(i)^T} \boldsymbol{h} y^{(i)} + \left(y^{(i)} \right)^2 \right] \\ &= \sum_{i=1}^{M} \left[\left(\boldsymbol{a}^{(i)^T} \boldsymbol{h} \right)^T \left(\boldsymbol{a}^{(i)^T} \boldsymbol{h} \right) - 2\boldsymbol{a}^{(i)^T} \boldsymbol{h} y^{(i)} + \left(y^{(i)} \right)^2 \right] \\ &= \sum_{i=1}^{M} \left[\boldsymbol{h}^T \left(\boldsymbol{a}^{(i)} \boldsymbol{a}^{(i)^T} \right) \boldsymbol{h} - 2y^{(i)} \boldsymbol{a}^{(i)^T} \boldsymbol{h} + \left(y^{(i)} \right)^2 \right] \\ &= \boldsymbol{h}^T \left(\sum_{i=1}^{M} \left(\boldsymbol{a}^{(i)} \boldsymbol{a}^{(i)^T} \right) \right) \boldsymbol{h} - 2 \left(\sum_{i=1}^{M} y^{(i)} \boldsymbol{a}^{(i)^T} \right) \boldsymbol{h} + \sum_{i=1}^{M} \left(y^{(i)} \right)^2 \end{aligned} \tag{8}$$

Here, we set $\mathbf{Q} = \sum_{i=1}^{M} \left(\boldsymbol{a}^{(i)} \boldsymbol{a}^{(i)^T} \right)$ and $\mathbf{P} = \left(\sum_{i=1}^{M} y^{(i)} \boldsymbol{a}^{(i)^T} \right)$, respectively. Therefore, Eq. (8) can be rewritten as

$$J(\boldsymbol{h}) = \boldsymbol{h}^T \boldsymbol{Q} \boldsymbol{h} - 2\boldsymbol{P}\boldsymbol{h} + \sum_{i=1}^{M} \left(y^{(i)} \right)^2 \tag{9}$$

To learn the $\boldsymbol{h}$ of $\boldsymbol{D}$, we minimize the cost function $J(\boldsymbol{h})$ as:

$$\boldsymbol{h}^* = \underset{\boldsymbol{h}}{\arg\min}\, J(\boldsymbol{h}) = \underset{\boldsymbol{h}}{\arg\min}\left(\boldsymbol{h}^T\boldsymbol{Q}\boldsymbol{h} - 2\boldsymbol{P}\boldsymbol{h} + \sum_{i=1}^{M}\left(y^{(i)}\right)^2\right)$$
$$= \underset{\boldsymbol{h}}{\arg\min}\left(\boldsymbol{h}^T\boldsymbol{Q}\boldsymbol{h} - 2\boldsymbol{P}\boldsymbol{h}\right) \quad (10)$$
$$s.t.\boldsymbol{h} \subset [0, 1]$$

Eq. (10) is a quadratic programming problem; in the present study, the trust-region-reflective algorithm was used to solve this problem. Furthermore, the initial classification score of the cluster centers can be estimated and added to the algorithm to accelerate the convergence. The initialization of $\boldsymbol{h}$ is noted as $\boldsymbol{h}^{(0)} = [h_1^{(0)}, \cdots, h_i^{(0)}, \cdots h_N^{(0)}]^T$, and $h_i^{(0)}$ can be estimated as

$$h_i^{(0)} = \frac{1}{O(G_i)} \sum_{\boldsymbol{x}^{(j)} \in G_i} y^{(j)}, \quad (11)$$

where $y^{(j)}$ is the classification score associated with training sample $\boldsymbol{x}^{(j)}$, and $O(G_i)$ is the size of G_i defined in Eq. (6).

Summary of LLRC

To clearly elucidate the concept of LLRC, we provided a pseudo-code for LLRC in Algorithm 1.

Algorithm 1. LLRC

```
Input: Training set $\mathbf{T} = \{\boldsymbol{x}^{(i)}, y^{(i)}\}_{i=1}^{M}$; A test sample $\boldsymbol{x}$.
Output: The classification score $y$ of $\boldsymbol{x}$.
Stage 1 Dictionary learning
        Partition $\{\boldsymbol{x}^{(i)}\}_{i=1}^{M}$ into $N$ sub-sets using $k$-means methods and
        choose cluster centers to construct dictionary $\boldsymbol{D}$.
Stage 2 Calculation of locally linear representation coefficients
        Reconstruct each $\boldsymbol{x}^{(i)}$ in $\mathbf{T}$ based on dictionary $\boldsymbol{D}$ by using the
        LAE method and calculate the coefficient vector $\boldsymbol{a}^{(i)}$.
Stage 3 Classification score learning
        Learn $\boldsymbol{h} = (h_1, h_2, \cdots, h_N)^T$ by solving the quadratic programming
        problem in Eq. (10).
Stage 4 Core of LLRC
        Reconstruct the input sample $\boldsymbol{x}$ based on dictionary $\boldsymbol{D}$ by using
        the LAE method and calculate the coefficient vector $\boldsymbol{a}$.
        Calculate the classification score $y$ of $\boldsymbol{x}$ based on the obtained $\boldsymbol{h}$
        according to Eq. (4).
```

Preprocessing

Considering the findings of (Eskildsen et al., 2012), we used the normalization of intensity and space to reduce the variance across the images and improve the accuracy of brain extraction. In the current study, the following similar preprocessing steps were performed on all images in our training and test datasets. First, the N3 algorithm (Sled et al., 1998) was utilized to remove bias field artifacts from the images. Second, intensities in the scans were normalized according to a two-step method (Nyul et al., 2000). Third, spatial normalization was performed by linear registration (by FLIRT[5] using cross correlation as cost) to the publicly available ICBM152 average (Mazziotta et al., 2001). Afterward, the images were aligned to a standard template space with an image size of 193 × 229 × 193 and a voxel size of 1 mm × 1 mm × 1 mm. Finally, intensities were normalized again by performing the following steps. Intensity values at 0.1% and 99.9% of the histogram of the voxels were calculated in an approximate brain mask. These two values were then used to scale the voxel intensities linearly in the range [0, 100].

Considering that all of the images were linearly aligned to a standard template space, we defined an initial mask, which provided the volume of interest of classification to reduce the computational costs of the proposed method. Like (Eskildsen et al., 2012), we define the initial mask as

$$CM = \cup_{i=1}^{C} V_i - \cap_{i=1}^{C} V_i, \quad (12)$$

where V_i is the ith brain mask in the training dataset and C is the total number of brain masks. The defined initial mask is the difference between the union and the intersection of all brain masks in the training dataset. The proposed classification method is performed in this initial mask region under the assumption that the training dataset represents all brain sizes after spatial normalization.

Feature extraction

A patch-based technique is used to extract image feature. The intensity values in a patch around a voxel v are obtained and rearranged as a feature vector. To obtain spatial information, we sampled the coordinates of v and considered these factors as sub-features. The coordinates were normalized to [0, 100] beforehand to ensure consistency with the intensity value. In our study, the patch with the size of $w \times w \times w$ was used. Therefore, the final feature $\boldsymbol{x}$ of v is $(w^3 + 3)$-dimensional.

Postprocessing

The resolution of the standard template space may be different from that of the original image space; thus, to void the lack of accuracy, the segmentations obtained in the standard template space were wrapped back to the native image space by using the inverse of the linear registration ("Preprocessing" section).

Multi-resolution framework

To improve robustness and reduce computational costs of the proposed method, we used a multi-resolution framework similar to that in a previous study (Eskildsen et al., 2012). For a multi-resolution framework with P levels, classification originated from the coarsest level L_{P-1} to the finest level L_0 (original resolution) to extract the brain. The classification scores for all voxels in coarser levels were upsampled to initialize the classification for a finer level by using trilinear interpolation method. At each level, a confidence threshold α is defined to determine the specific voxels that can be labeled directly using the propagated initial classification scores and the specific voxels that require further processing. In the current study, the voxels with initial classification scores $y < \alpha$ were labeled as non-brain tissue and voxels with $y > 1 - \alpha$ were labeled as brain tissue. The remaining voxels with $y \in [\alpha, 1 - \alpha]$ were fed into the proposed LLRC classifier for an accurate classification. This procedure is repeated until the resolution of the original level L_0 is reached.

We used three levels ($P = 3$) with corresponding voxel sizes of 1 mm × 1 mm × 1 mm, 2 mm × 2 mm × 2 mm, and 4 mm × 4 mm × 4 mm. The threshold values (α) were set to 0.2 at L_1 and L_2 and 0.5 at L_0.

Table 3
Summary of the parameter settings used in the proposed method.

Parameter	Description	Setting
k	Number of the neighbors of LAE	25 at level L_2, 45 at level L_1, and 50 at level L_0
N	Dictionary size	30,000
w	Patch size $w \times w \times w$	5
P	Levels of multi-resolution	3
α	Confidence threshold described in "Postprocessing" section	0.2 at levels L_1 and L_2, and 0.5 at level L_0

[5] http://www.fmrib.ox.ac.uk/fsl/

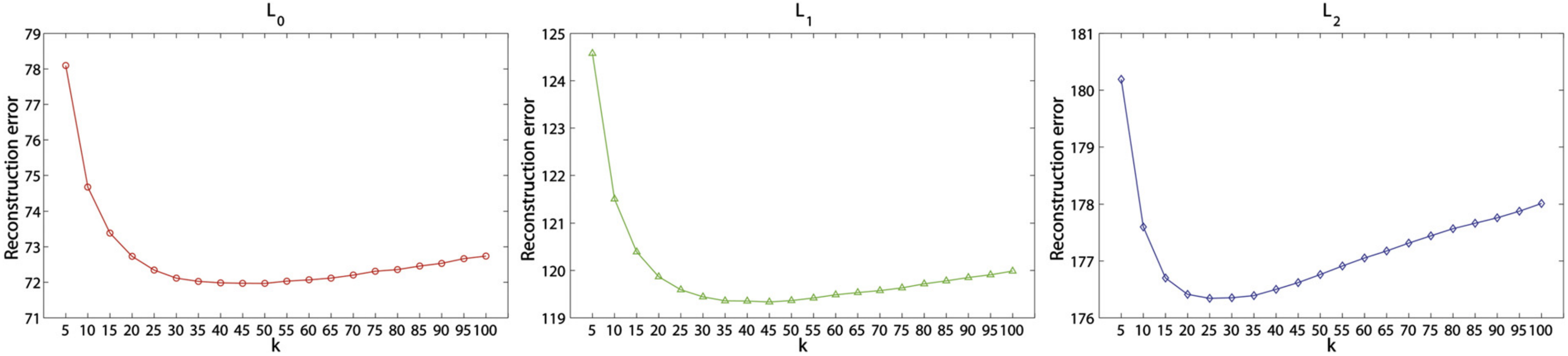

Fig. 2. Effect of different numbers of neighbors on the reconstruction error.

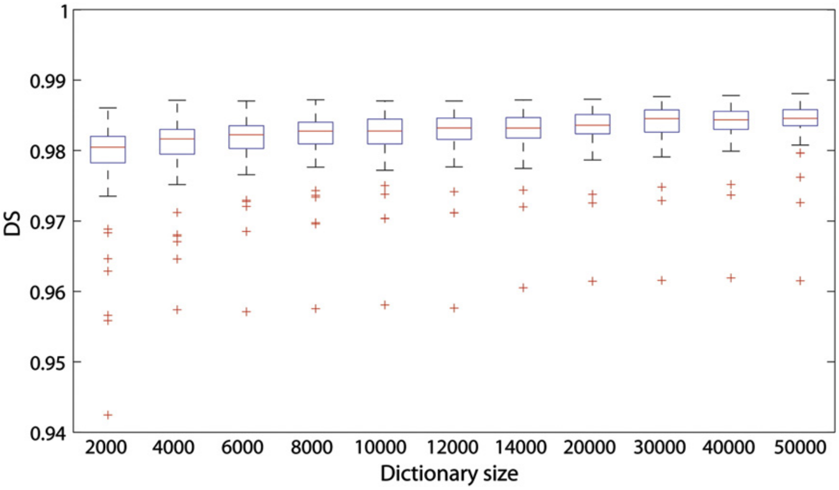

Fig. 3. Box-whisker plot of DS of classification using different dictionary sizes. Experiments performed by LOOCV with 140 images. The boxes indicate the 25th, 50th, and 75th percentiles. The whiskers indicate the most extreme data points excluding outliers (within three standard deviations from the mean). Outliers are marked as red crosses.

Experimental results

Assessment measures

To evaluate the performance of the algorithms, we used four quantitative metrics, namely, Dice similarity coefficient (DS), Jaccard similarity (JS), false positive rate (FPR), and false negative rate (FNR). These measures are defined as follows:

$$DS = \frac{2|A \cap B|}{|A| + |B|}, JS = \frac{|A \cap B|}{|A \cup B|}, FPR = \frac{|A \cap \overline{B}|}{|A \cup B|}, FNR = \frac{|\overline{A} \cap B|}{|A \cup B|}$$

where *A* and *B* are the voxel sets of the extraction result and brain mask, respectively. DS and JS can be used to measure the errors between the extraction result and brain mask; FPR and FNR can be used to quantify the extent of overextraction and underextraction.

In addition, the projection maps of false positive and false negative voxels were used to evaluate the algorithms qualitatively. In the normalized spatial space described in the "Preprocessing" section, the false positive (false negative) mask based on the false positive (false negative) error voxels of extraction can be obtained for each test image. All of these masks in the test database were averaged, and a mean false positive (false negative) mask image was obtained. The intensities of these mean mask images indicated the positions and frequencies of the error voxels of the test images. For simple visualization, the projection maps of the mean mask image on the sagittal, coronal, and axial directions were calculated and used for qualitative evaluation.

Parameter optimization

To optimize the parameters, a series of experiments was performed on the training dataset in a leave-one-out cross-validation (LOOCV) manner. Each of the 140 images in the training dataset was used as test image, and the remaining 138 images (aside from the test image and its flipped image) were used as training images. Under this situation, the proposed method was tested with different parameter settings to obtain the setting that achieves the best performance. The parameter settings used in the proposed method are summarized in Table 3.

A straightforward way was adopted to determine parameter *k* (defined in the "Locally linear representation" section), which represents the number of neighbors used by LAE. A number of random samples from the training dataset were reconstructed by LAE with varying *k*. The reconstructed errors were then used to evaluate the effectiveness of *k*. The *k* that results in the minimum reconstruction error is regarded as the optimum selection. In our experiment, the dictionary size was set to 30,000, and the patch size *w* was fixed to 5. A total of 20,000 samples were randomly sampled from the training dataset. This procedure was performed at each resolution level, and the result is shown in Fig. 2. The reconstruction errors reach their minimum when *k* is set to 25 at

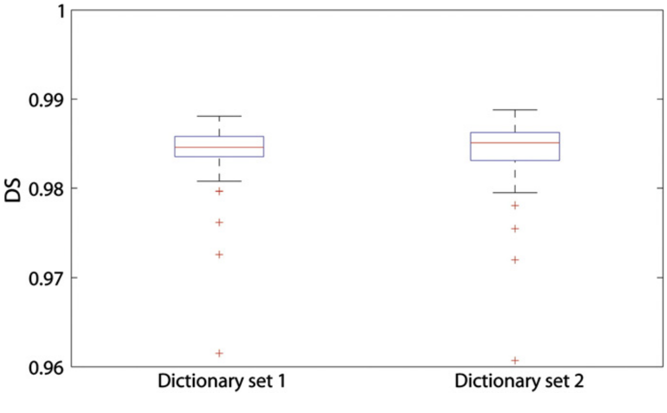

Fig. 4. DS of classification using different random initializations.

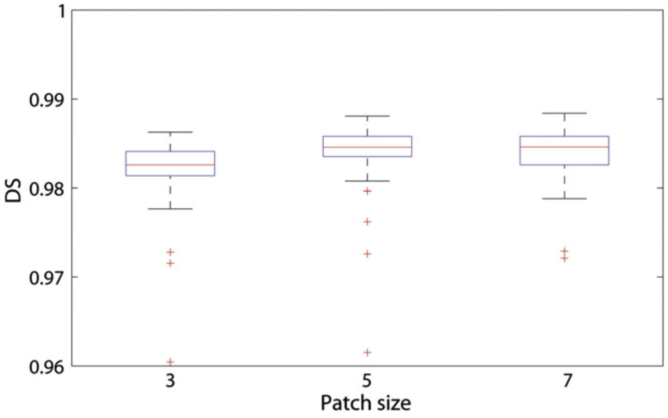

Fig. 5. DS of classification using different patch sizes.

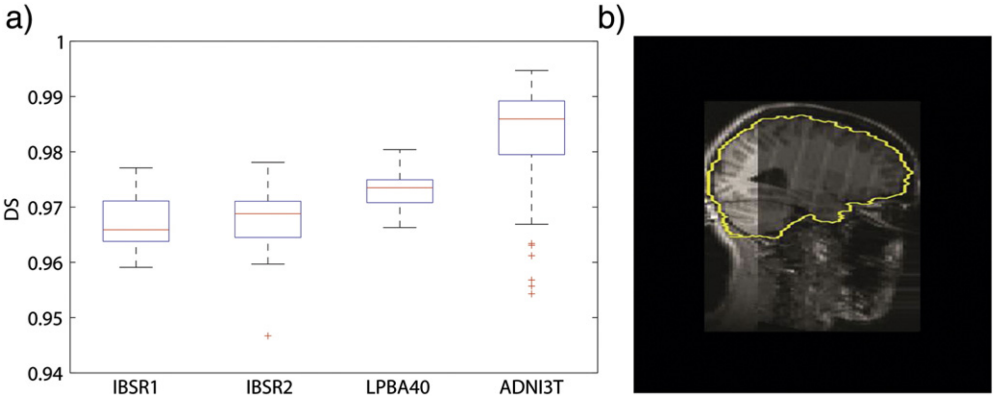

Fig. 6. (a) Dice similarity of the proposed method on different test datasets. (b) A slice of the outlier in IBSR2 dataset using the proposed method. The brain extraction result in this slice using the proposed method is shown in yellow curve.

level L_2, 45 at level L_1, and 50 at level L_0. Thus, the optimal selection of k has been performed for the subsequent experiments.

Dictionary learning is an important step of LLRC. As k-means was adopted to create the dictionary, two factors should be carefully considered in the implementation: the dictionary size N and the initialization of k-means. To optimize the dictionary size and investigate the influence of the initialization of k-means on the performance of LLRC, two experiments were conducted in an LOOCV manner. Patch size w was fixed to 5. Parameter k was set to 25 at level L_2, 45 at level L_1, and 50 at level L_0. For each level, a dictionary was constructed. Therefore, three dictionaries were created with the proposed method. The same N was assigned to these three dictionaries. In the first experiment, the proposed method was tested using different dictionaries with sizes varying from 2000 to 50,000. The result (see Fig. 3) shows that classification accuracy was improved by increasing N, in which the average DS increased from 0.9782 ($N = 2000$) to 0.9843 ($N = 50{,}000$) with gradually decreasing standard deviation. However, a larger dictionary corresponds to higher computation and memory costs. To attain a trade-off between the memory and computation costs and the accuracy, the dictionary size N was set to 30,000 in subsequent experiments. In the second experiment, k-means clustering was performed twice with different random initializations at each resolution level to create two sets of

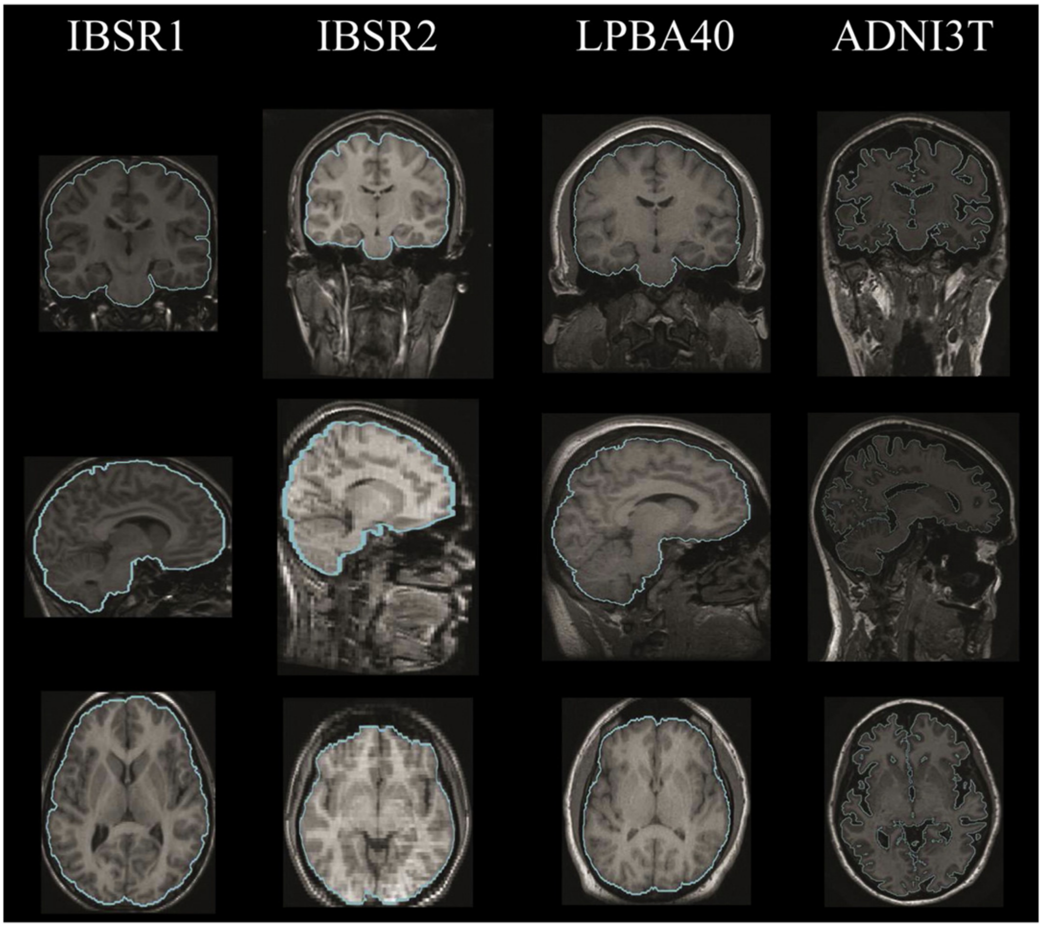

Fig. 7. Example of brain extraction results using the proposed method on each test dataset. From first to third column: coronal, sagittal, and axial views of each subject. The blue curve in each image indicates the outline of the brain extracted by the proposed method.

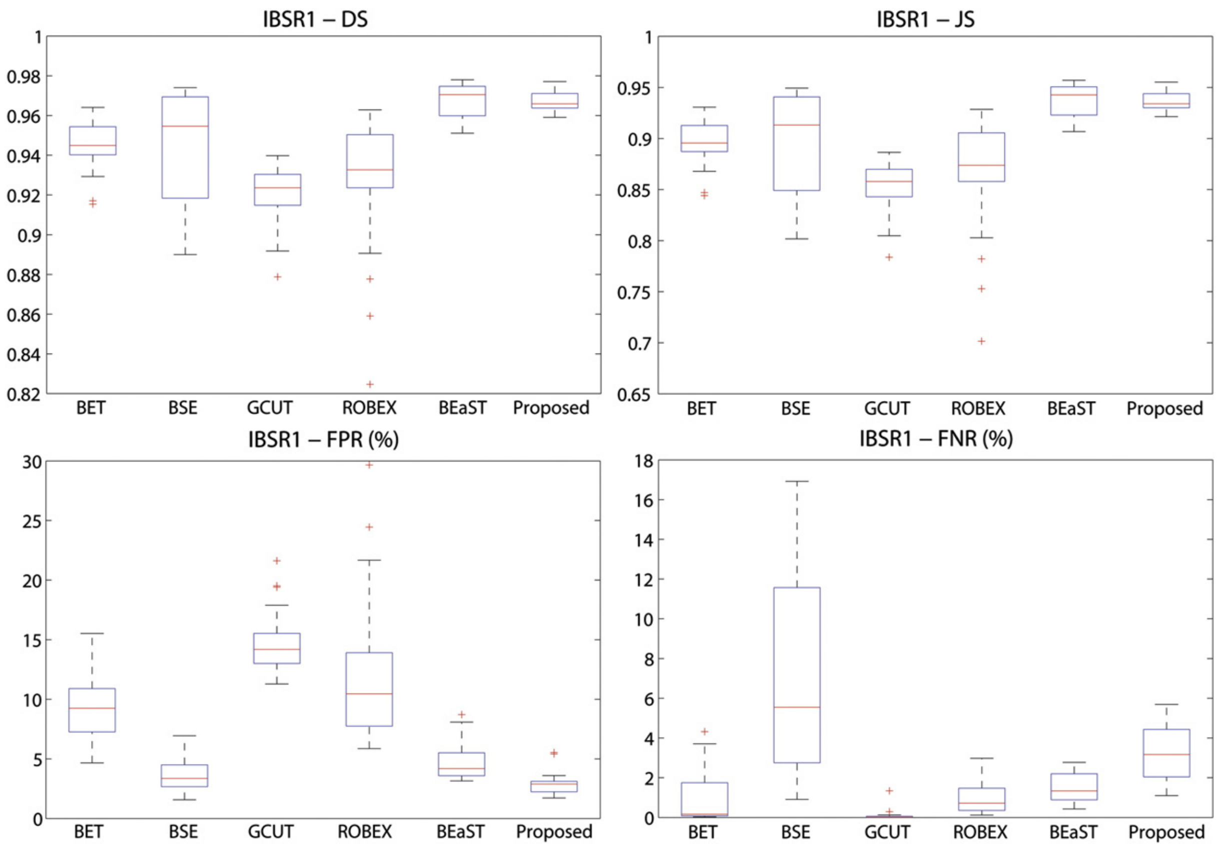

Fig. 8. Dice similarity, Jaccard similarity, false positive rate and false negative rate of different brain extraction methods in IBSR1 dataset.

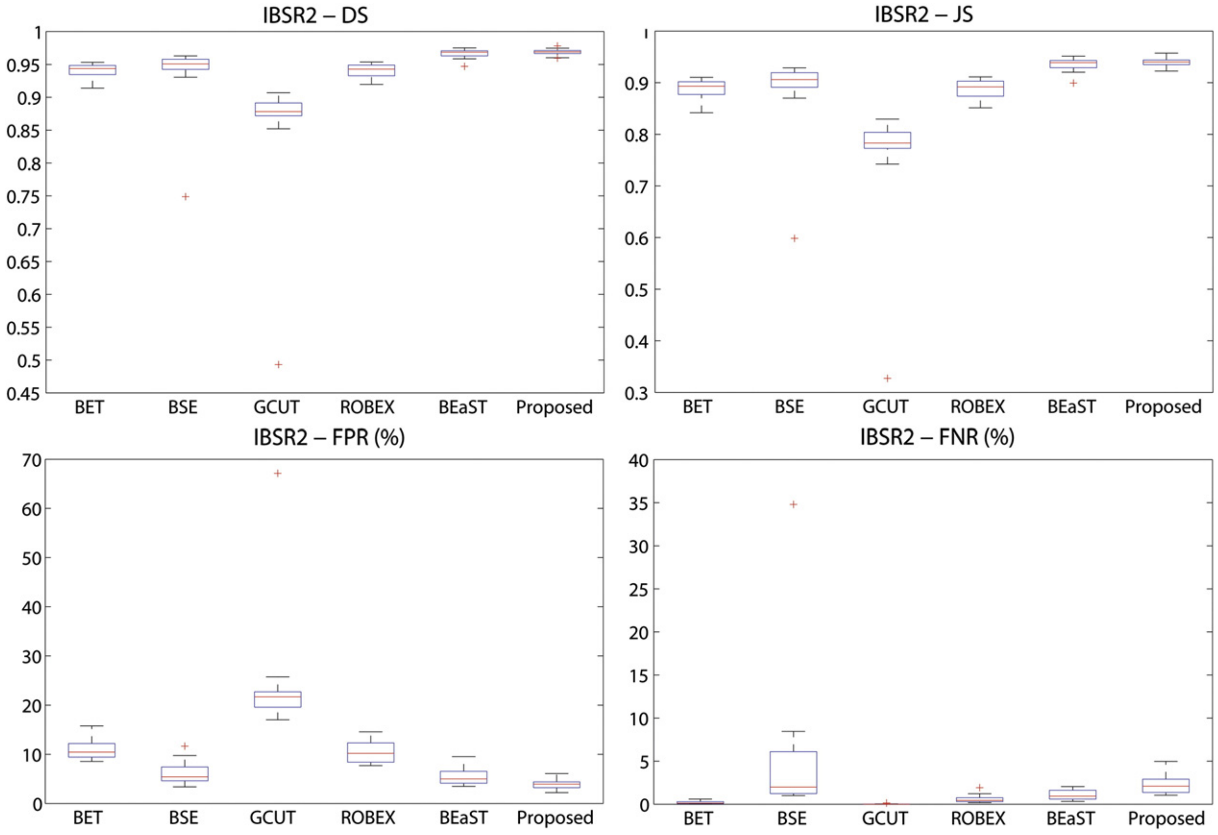

Fig. 9. Dice similarity, Jaccard similarity, false positive rate and false negative rate of different brain extraction methods in IBSR2 dataset. The outlier mentioned in "Evaluation on four test datasets" section in IBSR2 dataset has been excluded from the analysis.

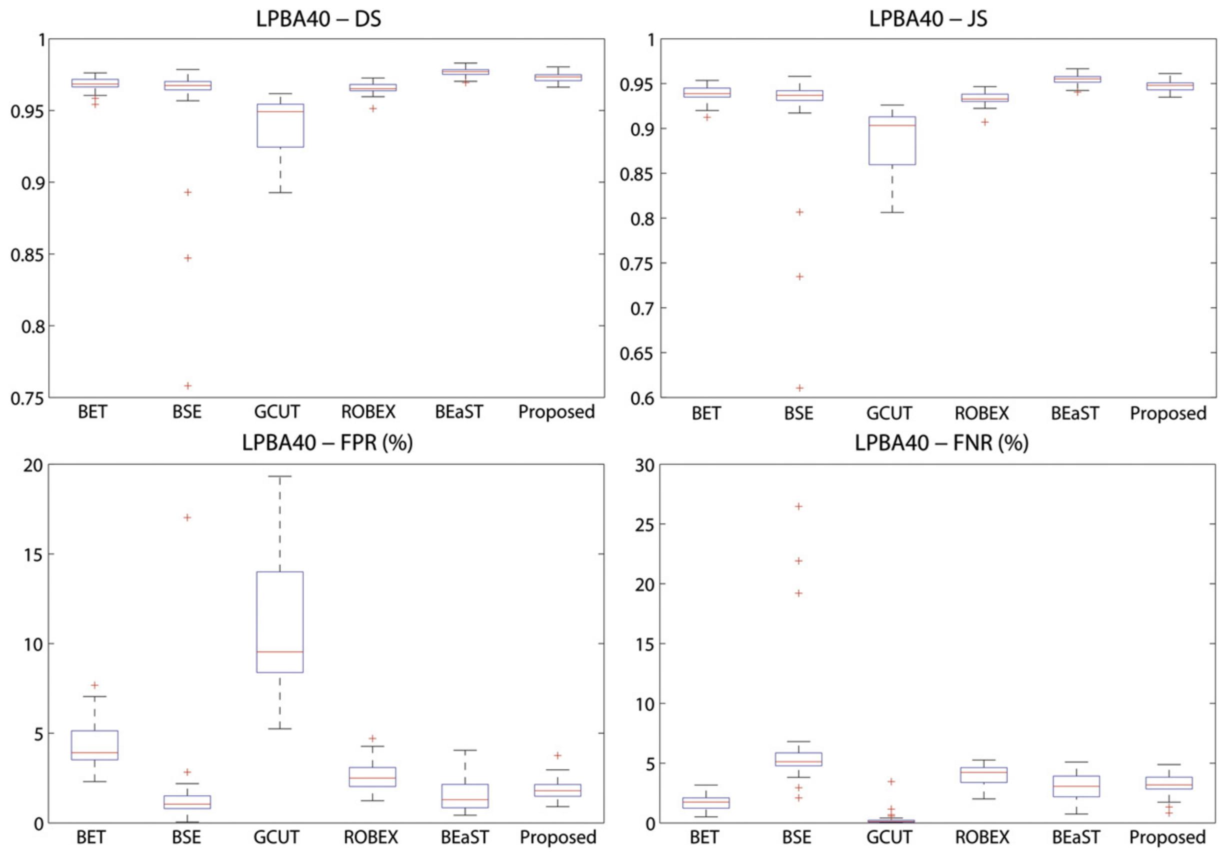

Fig. 10. Dice similarity, Jaccard similarity, false positive rate and false negative rate of different brain extraction methods in LPBA40 dataset.

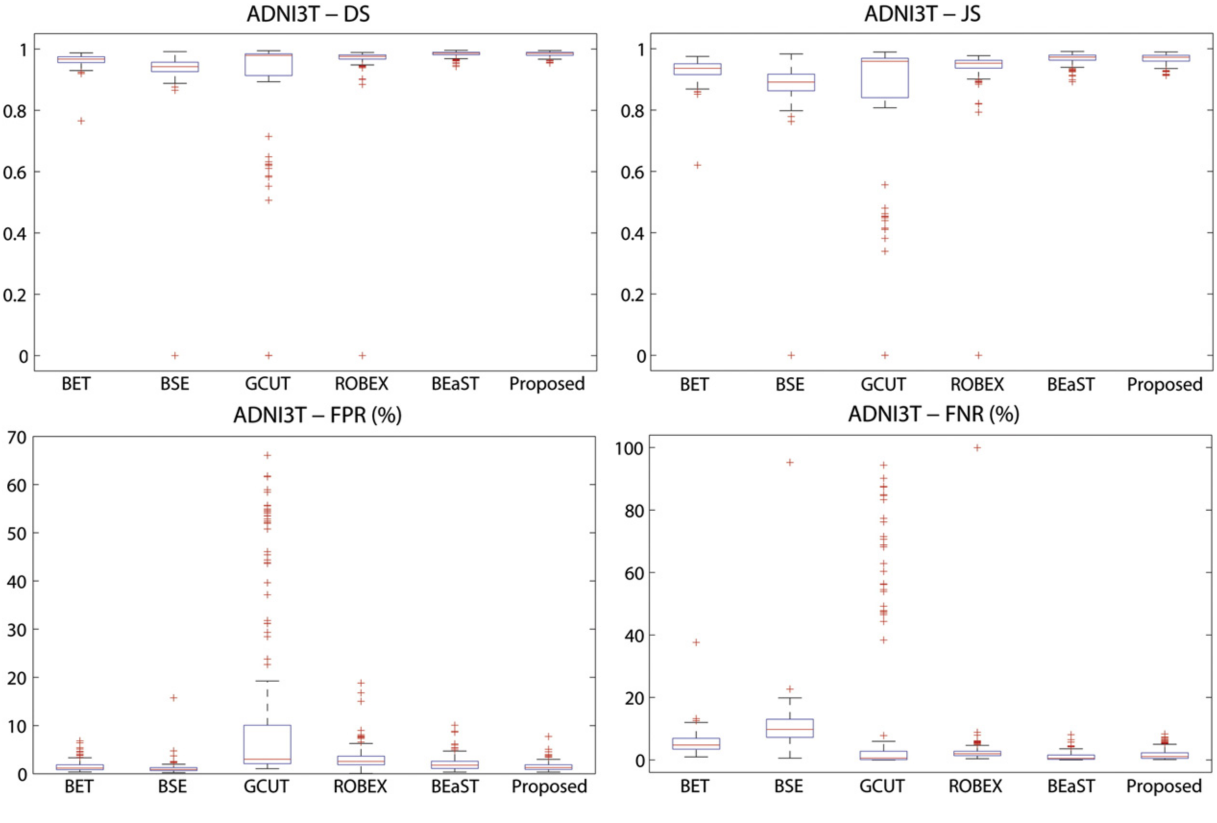

Fig. 11. Dice similarity, Jaccard similarity, false positive rate and false negative rate of different brain extraction methods in ADNI3T dataset.

Table 4
Mean ± standard deviations of the evaluation metrics with different brain extraction methods on the IBSR1 dataset. The best results from each column are shown in bold. *p*-Values of paired *t*-tests comparing the different methods with the proposed method are listed in the bottom five rows.

Method	DS	JS	FPR%	FNR%	DS (without CSF)	JS (without CSF)	FPR% (without CSF)	FNR% (without CSF)
BET	0.945 ± 0.014	0.896 ± 0.025	9.40 ± 2.90	1.03 ± 1.40	0.970 ± 0.013	0.941 ± 0.024	5.20 ± 2.77	0.69 ± 0.92
BSE	0.945 ± 0.028	0.896 ± 0.049	3.73 ± 1.58	6.67 ± 4.67	0.957 ± 0.019	0.918 ± 0.034	2.07 ± 1.31	6.15 ± 4.13
GCUT	0.919 ± 0.017	0.850 ± 0.029	14.9 ± 2.89	**0.12** ± **0.31**	0.968 ± 0.020	0.938 ± 0.037	6.07 ± 3.59	**0.12** ± **0.35**
ROBEX	0.925 ± 0.038	0.862 ± 0.063	12.9 ± 6.65	0.93 ± 0.76	0.962 ± 0.028	0.928 ± 0.051	6.43 ± 5.46	0.77 ± 0.54
BEaST	0.967 ± 0.009	0.937 ± 0.016	4.83 ± 1.73	1.43 ± 0.70	**0.983** ± **0.008**	**0.967** ± **0.016**	2.37 ± 1.69	0.97 ± 0.35
Proposed	**0.968** ± **0.006**	**0.937** ± **0.011**	**2.97** ± **1.04**	3.3 ± 1.52	0.979 ± 0.007	0.959 ± 0.014	**1.85** ± **1.13**	2.26 ± 0.98
Proposed method compared to other methods (p-values)								
BET	2.48×10^{-6}	2.19×10^{-6}	8.45×10^{-10}	2.11×10^{-6}	0.01	0.01	5.75×10^{-5}	3.33×10^{-5}
BSE	1.10×10^{-3}	1.00×10^{-3}	4.65×10^{-2}	3.40×10^{-3}	6.16×10^{-5}	4.98×10^{-5}	0.59	6.55×10^{-4}
GCUT	8.94×10^{-9}	5.63×10^{-9}	2.98×10^{-14}	1.47×10^{-7}	0.03	0.03	5.88×10^{-5}	1.11×10^{-9}
ROBEX	2.03×10^{-4}	1.20×10^{-4}	4.45×10^{-6}	1.35×10^{-7}	0.02	0.02	1.90×10^{-3}	4.63×10^{-6}
BEaST	0.95	0.99	1.57×10^{-6}	3.50×10^{-7}	0.15	0.14	0.29	1.52×10^{-5}

Table 5
Mean ± standard deviations of the evaluation metrics with different brain extraction methods on the IBSR2 dataset. The best results from each column are shown in bold. *p*-Values of paired *t*-tests comparing the different methods with the proposed method are listed in the bottom five rows. The outlier mentioned in the "Evaluation on four test datasets" section in IBSR2 dataset has been excluded from the analysis.

Method	DS	JS	FPR%	FNR%	DS (without CSF)	JS (without CSF)	FPR% (without CSF)	FNR% (without CSF)
BET	0.940 ± 0.011	0.887 ± 0.019	11.1 ± 2.05	0.20 ± 0.16	0.963 ± 0.013	0.929 ± 0.024	7.03 ± 2.48	0.12 ± 0.09
BSE	0.940 ± 0.047	0.890 ± 0.073	6.23 ± 2.14	4.76 ± 7.67	0.951 ± 0.047	0.909 ± 0.074	4.57 ± 2.28	4.53 ± 7.88
GCUT	0.861 ± 0.090	0.764 ± 0.109	23.62 ± 10.75	**0.002** ± **0.004**	0.935 ± 0.080	0.885 ± 0.112	11.46 ± 11.13	**0.02** ± **0.04**
ROBEX	0.942 ± 0.010	0.890 ± 0.017	10.44 ± 2.04	0.59 ± 0.44	0.967 ± 0.010	0.936 ± 0.019	5.85 ± 2.15	0.51 ± 0.34
BEaST	0.967 ± 0.007	0.936 ± 0.012	5.34 ± 1.60	1.07 ± 0.56	**0.974** ± **0.008**	**0.950** ± **0.015**	4.09 ± 1.75	0.85 ± 0.40
Proposed	**0.969** ± **0.005**	**0.939** ± **0.010**	**3.86** ± **1.01**	2.28 ± 1.10	0.972 ± 0.005	0.946 ± 0.009	**3.77** ± **1.07**	1.72 ± 0.86
Proposed method compared to other methods (p-values)								
BET	3.06×10^{-12}	1.98×10^{-12}	6.56×10^{-16}	1.05×10^{-9}	7.80×10^{-3}	8.10×10^{-3}	6.68×10^{-5}	1.34×10^{-9}
BSE	1.31×10^{-3}	6.20×10^{-3}	9.71×10^{-5}	0.17	0.03	0.02	0.19	0.13
GCUT	7.80×10^{-6}	2.71×10^{-8}	1.78×10^{-9}	1.07×10^{-10}	0.02	0.01	4.90×10^{-3}	3.05×10^{-10}
ROBEX	1.06×10^{-12}	7.80×10^{-13}	9.52×10^{-15}	3.80×10^{-7}	0.04	0.04	5.71×10^{-4}	1.83×10^{-6}
BEaST	0.37	0.39	1.60×10^{-3}	1.32×10^{-4}	0.20	0.20	0.49	3.28×10^{-4}

dictionaries, which were used to test the performance of classification. As shown in Fig. 4, the mean DS values of brain extraction were 0.9840 and 0.9836 for these two sets of dictionaries. This result indicates that the classification accuracy is not sensitive to the initialization of k-means (paired t-test $p = 0.5784$) when random selection approach is adopted.

Patch size w is a crucial parameter that should be carefully determined. In the current experiment, we also used the LOOCV method to evaluate the effect of w on classification performance. We set the same w at different levels and fixed w to 3, 5, and 7. In addition, parameter k was set to 25 at level L_2, 45 at level L_1, and 50 at level L_0 for different w. Fig. 5 shows that classification accuracy was improved (paired t-test $p = 0.006$) by increasing w from 3 to 5, in which the average DS increased from 0.9819 ($w = 3$) to 0.9836 ($w = 5$). However, the classification accuracy of $w = 7$ is similar to that of $w = 5$ (paired t-test $p = 0.9431$). Therefore, we chose $w = 5$ for subsequent experiments.

Table 6
Mean ± standard deviations of the evaluation metrics with different brain extraction methods on the LPBA40 dataset. The best results from each column are shown in bold. *p*-values of paired *t*-tests comparing the different methods with the proposed method are listed in the bottom five rows.

Method	DS	JS	FPR%	FNR%	DS (without CSF)	JS (without CSF)	FPR% (without CSF)	FNR% (without CSF)
BET	0.968 ± 0.005	0.939 ± 0.009	4.43 ± 1.26	1.72 ± 0.61	0.983 ± 0.006	0.966 ± 0.012	2.97 ± 1.23	0.45 ± 0.15
BSE	0.958 ± 0.039	0.921 ± 0.063	1.56 ± 2.57	6.36 ± 4.82	0.974 ± 0.035	0.951 ± 0.059	1.24 ± 1.52	3.68 ± 5.07
GCUT	0.939 ± 0.021	0.885 ± 0.036	11.3 ± 3.82	**0.26** ± **0.56**	0.970 ± 0.018	0.941 ± 0.034	5.72 ± 3.47	**0.15** ± **0.42**
ROBEX	0.966 ± 0.004	0.934 ± 0.007	2.58 ± 0.74	4.05 ± 0.89	0.981 ± 0.004	0.964 ± 0.007	1.47 ± 0.47	2.18 ± 0.62
BEaST	**0.977** ± **0.003**	**0.954** ± **0.006**	**1.48** ± **0.81**	3.10 ± 1.06	**0.989** ± **0.002**	**0.978** ± **0.005**	**1.09** ± **0.66**	1.06 ± 0.37
Proposed	0.973 ± 0.003	0.947 ± 0.006	1.87 ± 0.55	3.26 ± 0.93	0.986 ± 0.003	0.973 ± 0.005	1.45 ± 0.65	1.30 ± 0.38
Proposed method compared to other methods (p-values)								
BET	3.86×10^{-5}	4.82×10^{-5}	4.12×10^{-18}	9.85×10^{-20}	1.90×10^{-3}	1.90×10^{-3}	1.22×10^{-9}	1.93×10^{-21}
BSE	1.55×10^{-2}	1.03×10^{-2}	0.44	1.67×10^{-4}	0.03	0.02	0.43	4.10×10^{3}
GCUT	6.00×10^{-12}	2.87×10^{-12}	6.24×10^{-19}	5.08×10^{-23}	2.03×10^{-7}	1.52×10^{-7}	4.57×10^{-11}	5.36×10^{-21}
ROBEX	1.59×10^{-15}	1.92×10^{-15}	1.52×10^{-6}	2.36×10^{-6}	5.21×10^{-9}	5.13×10^{-9}	0.84	4.22×10^{-11}
BEaST	1.59×10^{-10}	5.43×10^{-11}	3.83×10^{-6}	0.07	1.56×10^{-6}	1.40×10^{-6}	0.02	5.30×10^{-3}

Table 7
Mean ± standard deviations of the evaluation metrics with different brain extraction methods on the ADNI3T dataset. The best results from each column are shown in bold. *p*-Values of paired *t*-tests comparing the different methods with the proposed method are listed in the bottom five rows.

Method	DS	JS	FPR%	FNR%
BET	0.963 ± 0.021	0.930 ± 0.035	1.54 ± 1.14	5.47 ± 3.61
BSE	0.936 ± 0.077	0.886 ± 0.081	**1.11** ± **1.29**	10.30 ± 7.88
GCUT	0.790 ± 0.364	0.762 ± 0.365	12.31 ± 18.47	11.54 ± 24.94
ROBEX	0.966 ± 0.077	0.939 ± 0.078	3.18 ± 2.49	2.88 ± 7.71
BEaST	0.983 ± 0.009	0.967 ± 0.017	2.17 ± 1.46	**1.10** ± **1.39**
Proposed	**0.984** ± **0.008**	**0.968** ± **0.016**	1.52 ± 0.96	1.72 ± 1.69
Proposed method compared to other methods (p-values)				
BET	1.67×10^{-26}	6.99×10^{-30}	0.84	3.54×10^{-28}
BSE	5.12×10^{-28}	1.00×10^{-30}	1.10×10^{-3}	3.77×10^{-34}
GCUT	3.70×10^{-11}	2.67×10^{-12}	5.32×10^{-13}	7.04×10^{-7}
ROBEX	3.20×10^{-3}	7.33×10^{-6}	1.82×10^{-14}	5.94×10^{-2}
BEaST	0.72	0.67	6.41×10^{-5}	2.75×10^{-4}

Performance of the proposed method

Evaluation on four test datasets

The proposed method was extensively evaluated on the IBSR1, IBSR2, LPBA40, and ADNI3T datasets. For ADNI3T, CSF should be excluded from the extracted brain before being compared with the masks because of its differences in mask definitions from the training datasets. As suggested by (Leung et al., 2011), a threshold of 60% of the mean intensity of the brain mask was used to exclude CSF from the segmented images in our experiment.

Fig. 6 (a) shows the DS of the classifications on these four test datasets. The classification accuracy on IBSR1 is similar to that on IBSR2 ($p = 0.9345$, two sample *t*-test). An outlier in IBSR2 may be caused by intensity heterogeneity and the registration error in the corresponding MRI image (Fig. 6 (b)). However, the intensity inhomogeneity of this outlier seems to be quite uncommon (Fig. 6 (b)). This outlier would be discarded by image quality control in a clinical study. Therefore, this outlier was excluded from subsequent analysis. In addition, the proposed method provided more accurate results on LPBA40 than on IBSR1 and IBSR2 ($p = 6.55 \times 10^{-5}$ and $p = 1.55 \times 10^{-4}$, respectively, two sample *t*-test). Furthermore, the brain extraction results of the proposed method achieved the highest scores on the ADNI3T dataset. Thus, the proposed method has reliability and robustness in working on these four datasets. Fig. 7 shows an example of the results of brain extraction with the use of the proposed method on each test dataset.

Computation and memory cost

In the current study, the experiments were implemented on a standard PC by using a single thread on an Intel Core i5-2400 processor at 3.10 GHz. In the brain extraction of a subject with $N = 30{,}000$ in the test step, the processing time was approximately 24 min, of which 4 min was allotted for intensity and spatial normalization, and 20 min was allotted for classification. In the training step, constructing a dictionary with $N = 30{,}000$ from a training dataset with 140 images consumed 5 h.

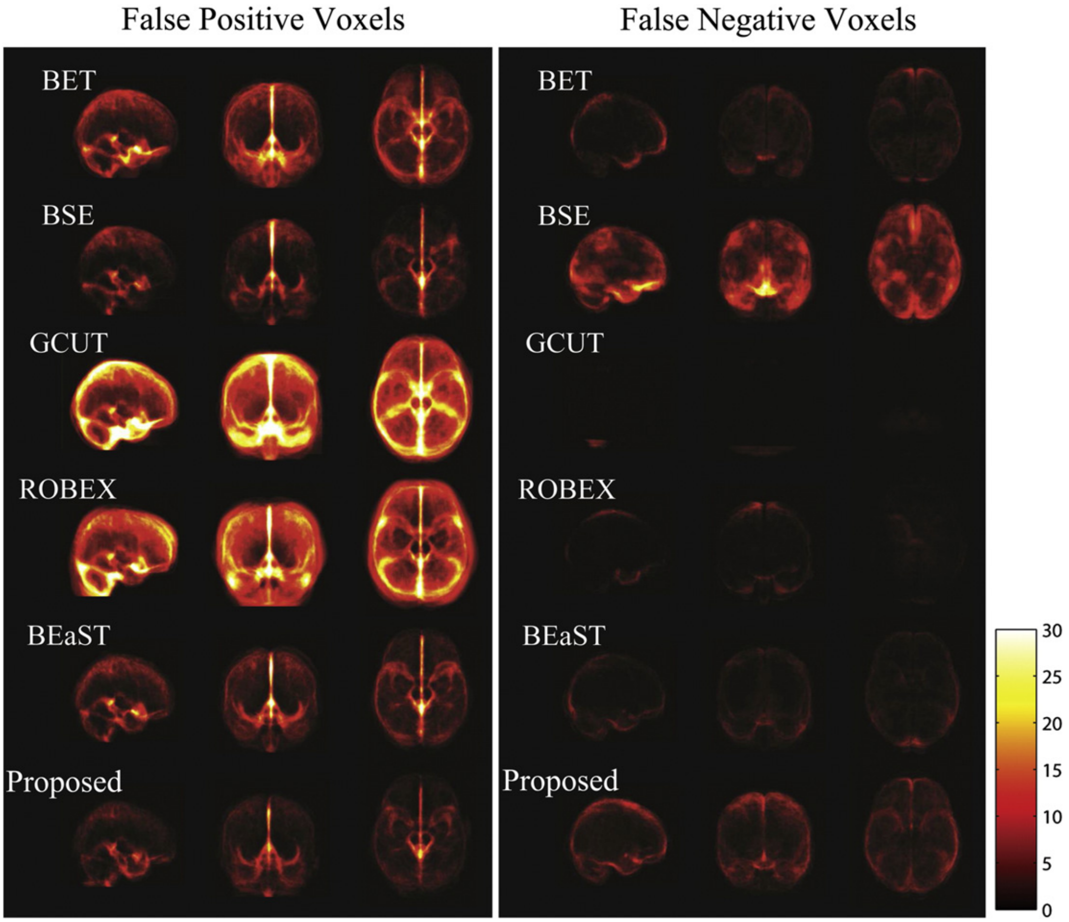

Fig. 12. Projection maps of false positives and false negatives for the brain extraction generated by BET, BSE, GCUT, ROBEX, BEaST, and the proposed method on IBSR1 dataset. All of the projection maps are shown in the same scale.

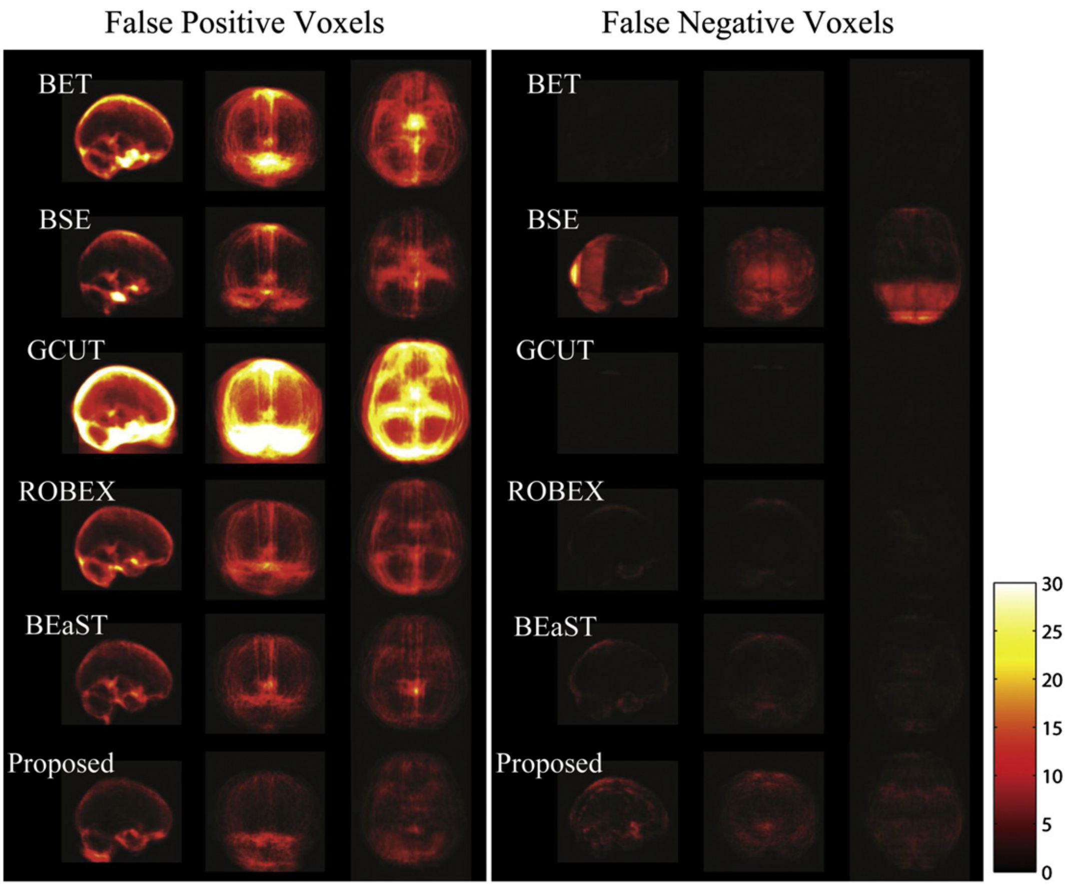

Fig. 13. Projection maps of false positives and false negatives for the brain extraction generated by BET, BSE, GCUT, ROBEX, BEaST, and the proposed method on IBSR2 dataset. All of the projection maps are shown in the same scale. The outlier mentioned in "Evaluation on four test datasets" section in IBSR2 dataset has been excluded from the analysis.

Memory cost is an important factor for the feasibility of a method. For the proposed method, the memory cost mainly includes the storage of the under-processing image and the dictionaries. In our implementation, three dictionaries with a size of 30,000 reside in the memory. In addition, the required memory is 128 × 30,000 × 3 × 4 = 43.9 megabytes (assuming one voxel needs 4 bytes to store).

Comparison with other methods

To evaluate the performance of the proposed method, we compared the proposed method with BET, BSE, GCUT, ROBEX, and BEaST. BET, BSE, and GCUT are widely used in several brain extraction comparisons (Iglesias et al., 2011; Shi et al., 2012). ROBEX and BEaST are two recently developed brain extraction methods. To conduct a rational comparison, we adjusted the parameters of each method until optimum performance was achieved. The succeeding sections describe the setups of these methods in detail.

Setups

BET (ver. 2.1, in FSL ver. 5.0.2). In the comparison experiment, BET was performed using intensity normalized and stereotaxically aligned images with default parameters (Eskildsen et al., 2012).

BSE (ver. 11a). For both IBSR1 and IBSR2 datasets, BSE was performed using the parameters as suggested in a previous study (Park and Lee, 2009). For volumes 7 to 12 in IBSR1 dataset, the following parameters were set: diffusion iterations = 3; diffusion constant = 25; edge constant = 0.6; and erosion size = 2. For other images in the IBSR1 dataset, the parameters were fixed as follows: diffusion iterations = 3; diffusion constant = 25; edge constant = 1; and erosion size = 2. For the IBSR2 dataset except for several images (7_8, 8_4, 13_3, 111_2, and 191_3), the following default parameters were used: diffusion iterations = 3; diffusion constant = 0.25; edge constant = 0.62; and erosion size = 1. The following parameters were used for the other images in the IBSR2 dataset: diffusion iterations = 3; diffusion constant = 5; edge constant = 0.75; and erosion size = 1. For the LPBA40 dataset, the following parameters were used as suggested in a previous study (Iglesias et al., 2011): diffusion iterations = 5; diffusion constant = 15; edge constant = 0.65; and erosion size = 1. For the ADNI3T dataset, the following parameters were used as suggested in a previous study (Leung et al., 2011): diffusion iterations = 4; diffusion constant = 20; edge constant = 0.7; and erosion size = 1.

GCUT. For the IBSR1, IBSR2, and LPBA40 datasets, the following default parameters were used as suggested by previous studies (Iglesias et al., 2011; Sadananthan et al., 2010): threshold = 36; importance of intensity = 2.3. For the ADNI3T dataset, we varied the threshold between 32 and 40 (with increments of 1) and the importance of intensity between 1 and 3 (with increments of 0.1). The best combinations were determined for the ADNI3T dataset.

ROBEX (ver. 1.0). All of the scans were oriented similar to the atlas before they were fed to ROBEX. ROBEX requires no parameter adjustments.

BEaST (ver. 1.15). All of the parameters were set as presented in a previous study (Eskildsen et al., 2012).

Comparison results

Figs. 8 to 11 show the box plots with different evaluated metrics for each brain extraction method and test dataset. Tables 4 to 7 list additional details of DS, JS, FPR, and FNR of all the methods in the four

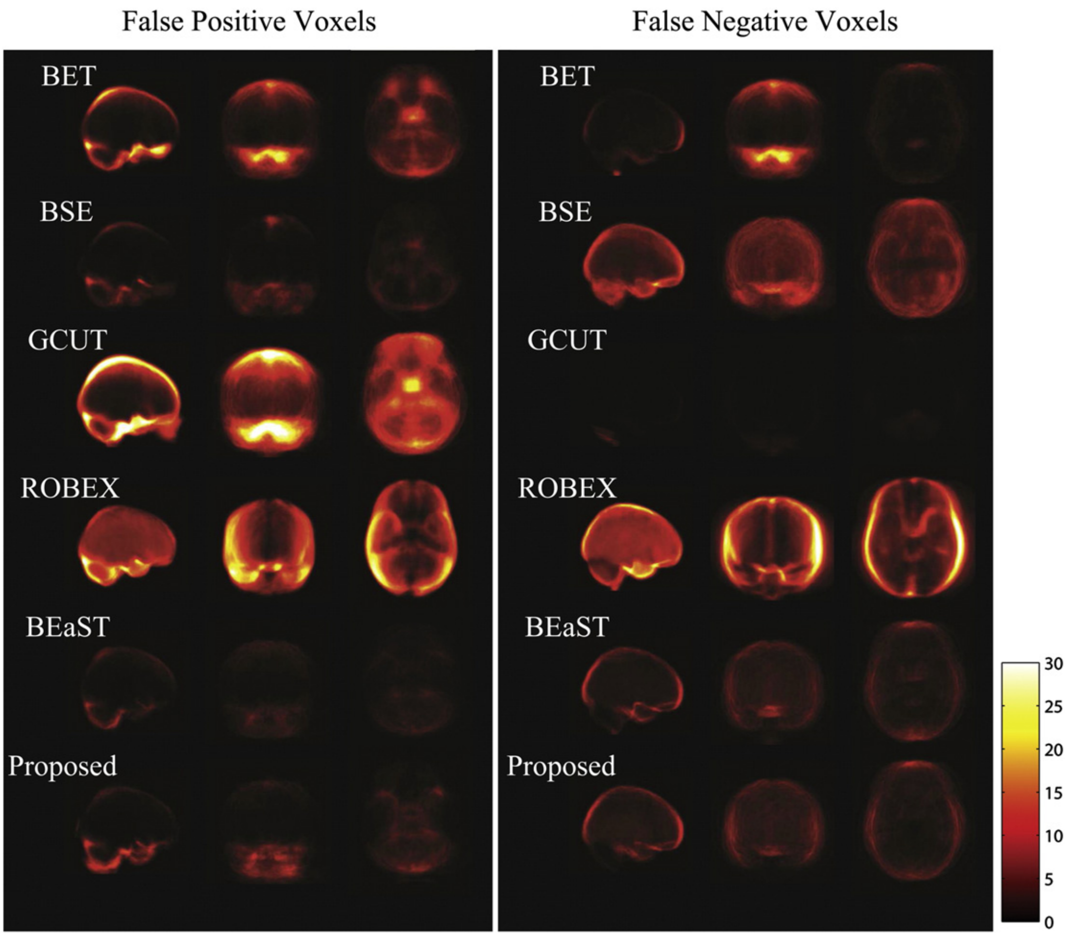

Fig. 14. Projection maps of false positives and false negatives for the brain extraction generated by BET, BSE, GCUT, ROBEX, BEaST, and the proposed method on LPBA40 dataset. All of the projection maps are shown in the same scale.

datasets. BET provided good results in the four datasets with the normalized data. The classification accuracy of BSE was similar to that of BET ($p > 0.1$, paired t-test) in the IBSR1, IBSR2, and LPBA40 datasets, but BSE produced more outliers than BET. BSE underextracted the images because of low FPR and high FNR; by contrast, BET overextracted the images with high FPR and low FNR. Moreover, BSE had a failed segmentation ($DS = 0$) in the ADNI3T dataset (Fig. 11). Although GCUT exhibited the lowest FNR (except for the ADNI3T dataset), FPR was high because GCUT maintained a large amount of non-brain tissues. Thus, GCUT obtained the lowest DS and JS values among the six methods. Moreover, GCUT performed well in most cases in the ADNI3T dataset but had 27 failed segmentations ($DS = 0$). ROBEX provided a consistent result in the IBSR2 dataset. However, ROBEX produced some outliers in the IBSR1 dataset, and exhibited high FPR in the IBSR1 and IBSR2 datasets. In addition, ROBEX had a failed segmentation ($DS = 0$) in the ADNI3T dataset. BEaST produced robust results in the four datasets. The proposed method yielded average DS values of 0.968 in the IBSR1 dataset, 0.969 in the IBSR2 dataset, 0.973 in the LPBA40 dataset, and 0.984 in the ADNI3T dataset. The proposed method has the highest extraction accuracy for the IBSR1, IBSR2, and ADNI3T datasets ($p < 0.01$ for all methods except BEaST; Tables 4, 5, and 7). FPR in the proposed method was also lower than that of BEaST in the IBSR1, IBSR2, and ADNI3T datasets. This result indicated that BEaST retained more non-brain tissues than the proposed method. However, FNR in the proposed method was higher than that of BEaST in the IBSR1, IBSR2, and ADNI3T datasets, indicating that the proposed method tends to remove more brain tissues than BEaST in these three datasets. The proposed method yielded the second highest DS values in the LPBA40 dataset after BEaST.

Figs. 12 to 15 show the projection maps for the false positive and false negative voxels to visualize the extraction errors in each brain extraction method and test dataset. Fig. 16 shows the typical results using different extraction methods in each dataset. BET likely retained extra non-brain tissues near the brain stem and on the top of the head (Figs. 12 to 14). BSE and BET behaved similarly in the IBSR2 dataset, but BSE removes more brain tissue than BET in the IBSR1, LPBA40, and ADNI3T datasets (Fig. 16). GCUT showed very few false negative voxels (except for the ADNI3T dataset), but a large amount of non-brain tissues were retained in the four datasets. ROBEX likely oversmoothened the brain surface, thereby leading to inclusion of the dura or exclusion of the GM (Fig. 16). BEaST included extra non-brain tissues near the bottom of the head in the IBSR1 and IBSR2 datasets (Figs. 12 to 13). The proposed method provided consistent and robust brain extractions in the four datasets.

The datasets have differences in their definitions of masks, and an arbitrary amount of CSF voxels may be included in the gold-standard skull-stripped images; therefore, segmentations that exclude the CSF were compared in our experiment. For the IBSR1, IBSR2, and LPBA40 datasets, a threshold of 36% of the mean intensity of WM was used to exclude CSF from the segmented images and the masks for fair comparison, as suggested by (Sadananthan et al., 2010). The results are listed in the last four columns of Tables 4 to 6. The brain extraction accuracy of the different methods without CSF is consistently higher than that with CSF. In addition, the performance of GCUT without CSF significantly improved (paired t-test p-value < 0.01) probably because GCUT intends to retain more non-brain tissues than the five other brain extraction methods.

Discussion

In the present study, a novel brain extraction method was proposed, in which the following advantages were observed. An LLRC method was developed to solve the brain extraction problem. Two assumptions were considered to elucidate the mechanism on why the labels of test samples can be estimated by linear combination of the training sample's labels in label fusion method. The LAE method was more applicable in solving the locally linear coefficients under locality constraint compared with other linear representation approaches. In addition, LLRC supplies a way to learn the optimal classification scores of the training samples in the dictionary.

This study is motivated by LLE (Roweis and Saul, 2000), which assumes that samples lie on a non-linear manifold and can be approximated locally and linearly. The weights of this linear approximation can be determined using several approaches, such as LLE (Roweis and Saul, 2000), SC (Wright et al., 2009), non-local means (Buades et al., 2005), LLC (Wang et al., 2010), and LAE (Liu et al., 2010). For dimensionality reduction, LLE calculates the weights by solving the linear least squares problem with a constraint that fixes the sum of the weights to 1. LLC adopts an approach similar to LLE to determine the weights; however, LLC is used for classification. SC emphasizes the sparsity of the representation, in which the lowest number of bases is used to reconstruct the target by means of basis pursuit approach. Non-local means determines the weights according to the distance between the samples by using Gaussian kernels, which cannot ensure the minimum reconstruction error but seems to be fit for denoising methods (Gong et al., 2010). In LAE, weight coefficients are maintained as non-negative values, and the sum is equal to one, which is a strong locality constraint and is desirable for LLRC to ensure the rationality of the Assumption 2 suggested in the current study.

In the present study, k-means method was used to generate the dictionary. In our experiment, we found dictionary size (i.e., the number of clusters) to be an important parameter that affects the performance of the classification. The experimental result shows that a larger dictionary corresponds to higher classification accuracy and greater computation and memory costs. An optimal parameter selection is achieved by making a trade-off between the memory and computation costs and the accuracy. In addition, the classification performance is not sensitive to the initialization of k-means when random selection approach is adopted. In future studies, other complex clustering methods can be used to improve performance further. For instance, we can learn multi-dictionaries for different spatial regions to capture more discriminative information on the objects or use a learning strategy to learn a dictionary, as reported in (Wang et al., 2010).

In the proposed method, patch size w was adjusted to optimize performance. In general, larger patches lead to more discriminative information for identifying different tissues. However, in high-dimensional feature spaces associated with larger patches, a larger dictionary is needed to construct the basis of the space, and more samples are required to train the model. In our experiment, the classification accuracy of $w = 5$ is similar to that of $w = 7$ (paired t-test $p = 0.9431$). Therefore, a moderate-sized patch was more applicable in the proposed method than patches of other sizes.

In the present study, six datasets were used to evaluate the performance of the proposed method. The classification accuracy of the LPBA40 dataset is higher than that of the IBSR1 and the IBSR2 datasets (Fig. 6). The enhanced performance of LPBA40 was caused by two reasons. First, the brain mask definition of the LPBA40 dataset is similar to that of the training dataset, whereas the CSF in the cerebellar cistern is excluded from the IBSR1 and IBSR2 test datasets. Second, LPBA40 exhibits good resemblance to the training data.

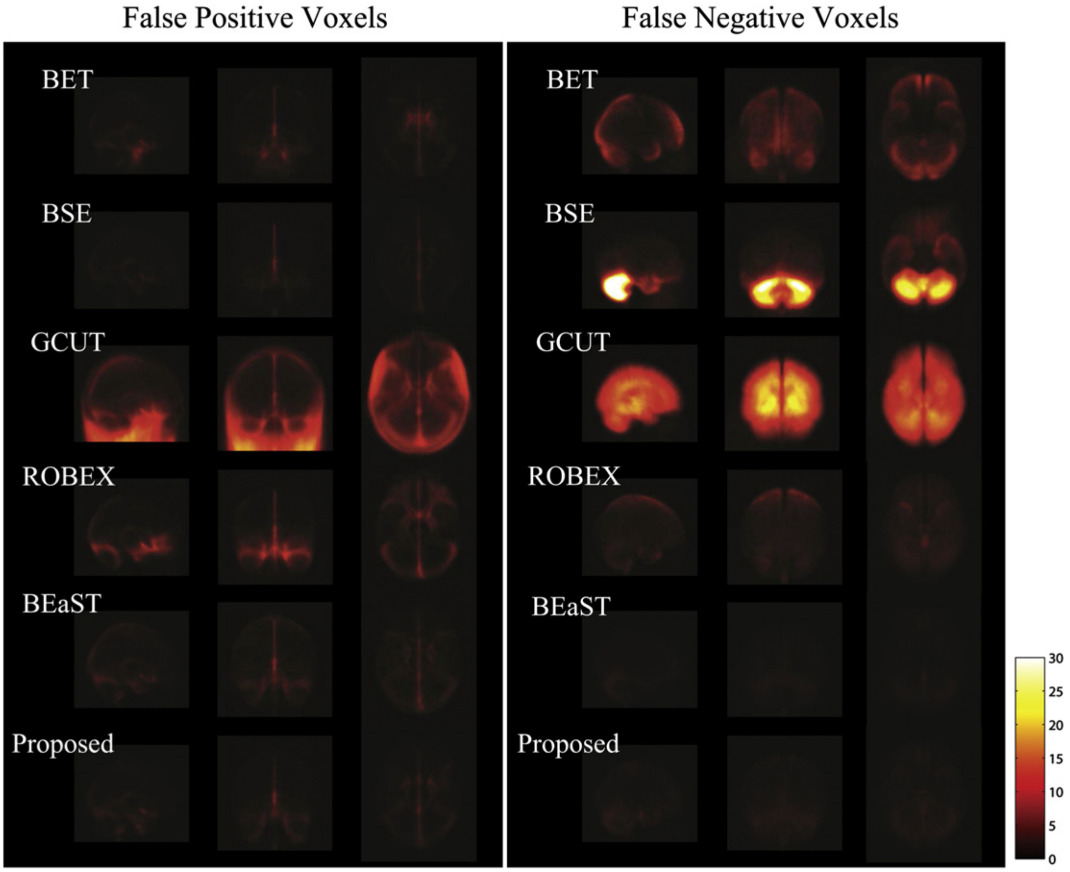

Fig. 15. Projection maps of false positives and false negatives for the brain extraction generated by BET, BSE, GCUT, ROBEX, BEaST, and the proposed method on ADNI3T dataset. All of the projection maps are shown in the same scale.

We compared the proposed method with other common brain extraction methods (BET, BSE, GCUT, ROBEX, and BEaST). Our performance evaluation results are consistent with those in previous findings (Iglesias et al., 2011; Park and Lee, 2009; Sadananthan et al., 2010). BET and BSE showed similar extraction accuracies, and the performances of these two methods are consistently high on the four test datasets. However, BET performed efficiently on normalized images with little or no visible neck; the performance of BSE was dependent on accurately adjusted parameters as previously reported (Iglesias et al., 2011). In GCUT, intensity thresholding is performed and narrow connections are then removed using the graph cut method, which may fail to cut connections without noticeable intensity separation between brain and non-brain tissues (Sadananthan et al., 2010), such as those in several volumes in the IBSR1 dataset. In ROBEX, the same problem was encountered, although a generative model can maintain the integrity of the brain extraction result. Thus, the least accurate brain extraction results of ROBEX were observed in the IBSR1 dataset. The proposed method provided a comparable performance to

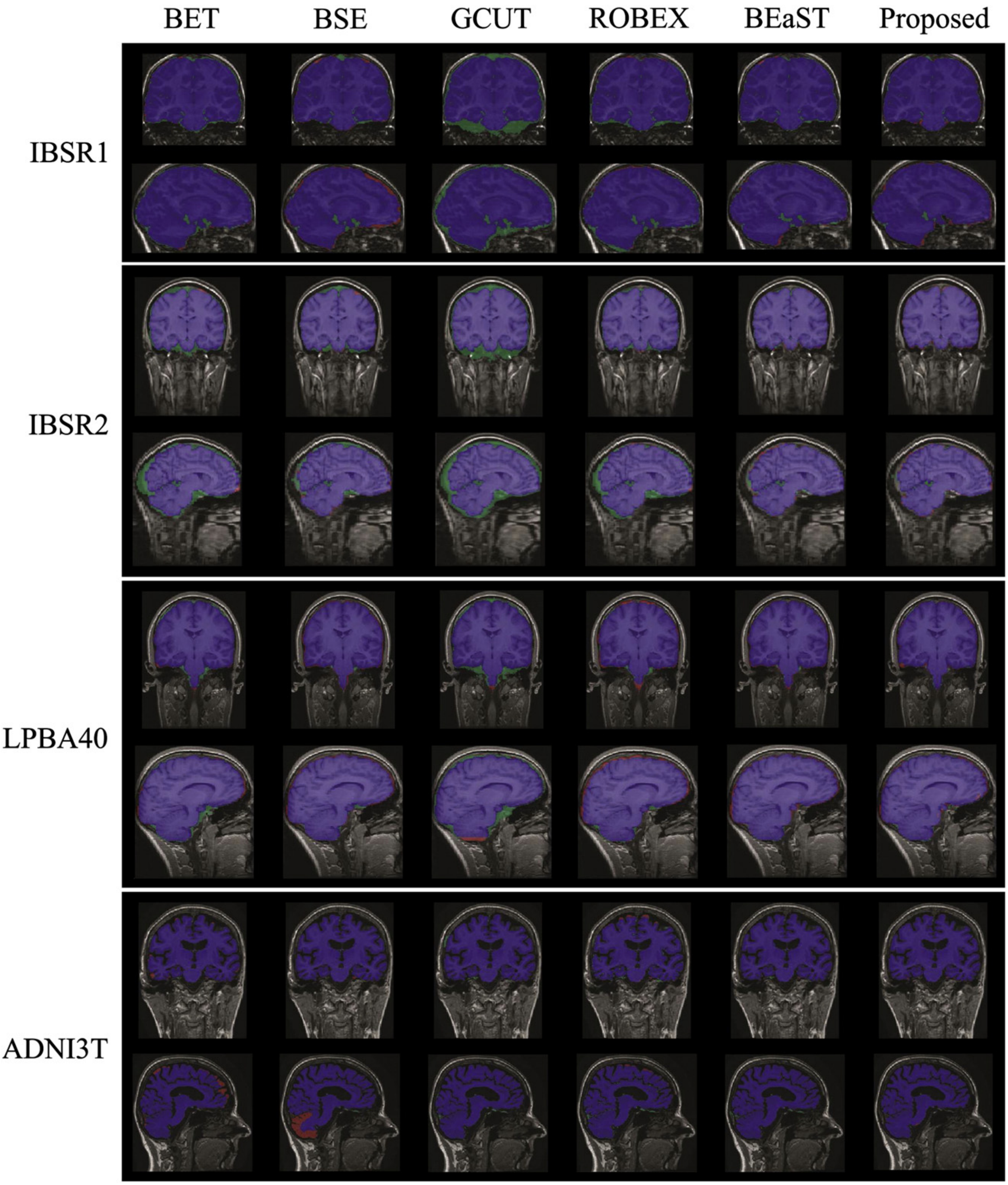

Fig. 16. Typical results using BET, BSE, GCUT, ROBEX, BEaST, and the proposed method on four datasets. Coronal and sagittal slices of the segmentations are shown. Blue voxels represent the segmentation results of the corresponding brain extraction methods. Green voxels indicate the false positives and red voxels indicate the false negatives.

BEaST on the IBSR1, IBSR2, and ADNI3T datasets, and highly outperforms other methods ($p < 0.05$) on the four test datasets.

In BEaST, several closest images from the library are first selected on the basis of their similarity to the target image. Label fusion is then performed for each voxel in the target image by using fusion weights to combine similar patches from the closest images. This label fusion method is the same as selecting samples from a very large dictionary that consists of all the samples in the training dataset. As previously mentioned, a large dictionary generally benefits classification accuracy. This is an important factor for BEaST to achieve very high accuracy. However, a large dictionary increases memory cost. In our experiment, the training dataset includes 140 images with a size of $193 \times 229 \times 193$. To perform BEaST, all these images need to be loaded into the memory, whose size should at least be $193 \times 229 \times 193 \times 140 \times 4 = 4.45$ gigabytes (assuming one voxel needs 4 bytes to store). Compared with that of BEaST, the memory cost of the proposed method is much lower ($128 \times 30000 \times 3 \times 4 = 43.9$ megabytes, "Computation and memory cost" section). In other words, the proposed method achieves competitive accuracy with significantly lower memory cost ($\frac{128\times30000\times3\times4}{193\times229\times193\times140\times4} = \frac{1}{104}$) than BEaST. This property is an attractive advantage of the proposed method. In addition to the dictionary size, the spatial constraint of the patch searching scheme in BEaST contributes to the high accuracy. In BEaST, similar patches are selected within a search area, which is a box centered at the testing voxel in the template space. Therefore, the selected patches are very likely to be sampled from the same structure with the testing patch. There is a strong correlation between the selected patches and the testing patch, which is beneficial to the classification performance.

We used only the patch and coordinate features, which may insufficiently discriminate the brain extraction task because of the complex characteristics of the brain MRI images. In future studies, other appearance and context features could be added to further improve the classification accuracy obtained in the present study. Although we focused only on T1-weighted MRI images in the current study, we can also apply the proposed method in other modalities, particularly if the locally linear representation-based classifier is trained with data acquired using the modality. If images from more than one modality are available, features from the images of the provided modality can be extracted; as such, classification results will possibly be improved.

Computation time is another aspect of the proposed method that could be improved. The classification step consumes approximately 20 min in the proposed method at $N = 30{,}000$. This implementation was run in a single thread. However, the voxel-based classification in the current study can be parallelized and implemented to use multi-core CPUs, thereby significantly decreasing processing time.

Brain extraction is an important pre-processing step in brain image analysis. Therefore, the effect of a brain extraction method on the subsequent analysis, such as cortical thickness measurement and brain atrophy estimation, is also a crucial factor to evaluate the performance of the brain extraction method. However, the effect of the proposed method on the subsequent analysis is not evaluated in the current study. This is a limitation in the current study.

In conclusion, this study presented a novel brain extraction method based on LLRC within a multi-resolution framework. The proposed method was compared with BET, BSE, GCUT, ROBEX, and BEaST in terms of four datasets. The accuracy of the proposed method is higher than that of BET, BSE, GCUT, and ROBEX, and comparable to that of BEaST.

Acknowledgments

This work was supported by grants from the National Basic Research Program of China (973 Program) (No. 2010CB732505), the National Science & Technology Pillar Program of China (No. 2012BAI14B02), the National Natural Science Funds of China (NSFC, No. 81101109 and No. 31371009), the National High Technology Research and Development Program of China (863 Program) (No. 2012AA02A616), and the Program of Pearl River Young Talents of Science and Technology in Guangzhou (No. 2013J2200065).

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 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 namely 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 (http://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 disseminated 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.

References

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry—the methods. NeuroImage 11, 805–821.

Baillard, C., Hellier, P., Barillot, C., 2001. Segmentation of brain 3D MR images using level sets and dense registration. Med. Image Anal. 5, 185–194.

Buades, A., Coll, B., Morel, J.M., 2005. A review of image denoising algorithms, with a new one. Multiscale Model. Simul. 4, 490–530.

Carass, A., Cuzzocreo, J., Wheeler, M.B., Bazin, P.L., Resnick, S.M., Prince, J.L., 2011. Simple paradigm for extra-cerebral tissue removal: algorithm and analysis. NeuroImage 56, 1982–1992.

Chiverton, J., Wells, K., Lewis, E., Chen, C., Podda, B., Johnson, D., 2007. Statistical morphological skull stripping of adult and infant MRI data. Comput. Biol. Med. 37, 342–357.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis. I. Segmentation and surface reconstruction. NeuroImage 9, 179–194.

Eskildsen, S.F., Coupe, P., Fonov, V., Manjon, J.V., Leung, K.K., Guizard, N., Wassef, S.N., Ostergaard, L.R., Collins, D.L., Alzheimer's Disease Neuroimaging I., 2012. BEaST: brain extraction based on nonlocal segmentation technique. NeuroImage 59, 2362–2373.

Gao, Y., Liao, S., Shen, D., 2012. Prostate segmentation by sparse representation based classification. Med. Phys. 39, 6372–6387.

Gong, D., Sha, F., Medioni, G., 2010. Locally linear denoising on image manifolds. 13th International Conference on Artifical Intelligence and Statistics.

Hahn, H.K., Peitgen, H.-O., 2000. The skull stripping problem in MRI solved by a single 3D watershed transform. Med. Image Comput. Comput. Assist. Interv. 134–143.

Iglesias, J.E., Liu, C.Y., Thompson, P.M., Tu, Z.W., 2011. Robust brain extraction across datasets and comparison with publicly available methods. IEEE Trans. Med. Imaging 30, 1617–1634.

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

Lemieux, L., Hagemann, G., Krakow, K., Woermann, F.G., 1999. Fast, accurate, and reproducible automatic segmentation of the brain in T1-weighted volume MRI data. Magn. Reson. Med. 42, 127–135.

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, 1091–1108.

Liu, W., He, J.F., Chang, S.-F., 2010. Large graph construction for scalable semi-supervised learning. 27th International Conference on Machine Learning.

Macqueen, J., 1967. Some methods for classification and analysis of multivariate observations. Fifth Berkeley Symp. on Math. Statist. and Prob, pp. 281–297.

Mazziotta, J.C., Toga, A.W., Evans, A., Fox, P., Lancaster, J., 1995. A probabilistic atlas of the human brain: theory and rationale for its development. The International Consortium for Brain Mapping (ICBM). NeuroImage 2, 89–101.

Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., Woods, R., Paus, T., Simpson, G., Pike, B., Holmes, C., Collins, L., Thompson, P., MacDonald, D., Iacoboni, M., Schormann, T., Amunts, K., Palomero-Gallagher, N., Geyer, S., Parsons, L., Narr, K., Kabani, N., Le Goualher, G., Boomsma, D., Cannon, T., Kawashima, R., Mazoyer, B.,

2001. A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). Philos. Trans. R. Soc. Lond. B Biol. Sci. 356, 1293–1322.

Mikheev, A., Nevsky, G., Govindan, S., Grossman, R., Rusinek, H., 2008. Fully automatic segmentation of the brain from T1-weighted MRI using bridge burner algorithm. J. Magn. Reson. Imaging 27, 1235–1241.

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

Nyul, L.G., Udupa, J.K., Zhang, X., 2000. New variants of a method of MRI scale standardization. IEEE Trans. Med. Imaging 19, 143–150.

Park, J.G., Lee, C., 2009. Skull stripping based on region growing for magnetic resonance brain images. NeuroImage 47, 1394–1407.

Rehm, K., Schaper, K., Anderson, J., Woods, R., Stoltzner, S., Rottenberg, D., 2004. Putting our heads together: a consensus approach to brain/non-brain segmentation in T1-weighted MR volumes. NeuroImage 22, 1262–1270.

Rex, D.E., Shattuck, D.W., Woods, R.P., Narr, K.L., Luders, E., Rehm, K., Stoltzner, S.E., Rottenberg, D.A., Toga, A.W., 2004. A meta-algorithm for brain extraction in MRI. NeuroImage 23, 625–637.

Roweis, S.T., Saul, L.K., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290, 2323–2326.

Sadananthan, S.A., Zheng, W.L., Chee, M.W.L., Zagorodnov, V., 2010. Skull stripping using graph cuts. NeuroImage 49, 225–239.

Segonne, F., Dale, A.M., Busa, E., Glessner, M., Salat, D., Hahn, H.K., Fischl, B., 2004. A hybrid approach to the skull stripping problem in MRI. NeuroImage 22, 1060–1075.

Shan, Z.Y., Yue, G.H., Liu, J.Z., 2002. Automated histogram-based brain segmentation in T1-weighted three-dimensional magnetic resonance head images. NeuroImage 17, 1587–1598.

Shattuck, D.W., Sandor-Leahy, S.R., Schaper, K.A., Rottenberg, D.A., Leahy, R.M., 2001. Magnetic resonance image tissue classification using a partial volume model. NeuroImage 13, 856–876.

Shen, D., Davatzikos, C., 2004. Measuring temporal morphological changes robustly in brain MR images via 4-dimensional template warping. NeuroImage 21, 1508–1517.

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

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging 17, 87–97.

Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17, 143–155.

van der Kouwe, A.J., Benner, T., Salat, D.H., Fischl, B., 2008. Brain morphometry with multiecho MPRAGE. NeuroImage 40, 559–569.

Wang, J.J., Yang, J.C., Yu, K., Lv, F.J., Huang, T., Gong, Y.H., 2010. Locality-constrained linear coding for image classification. Comput. Vis. Pattern Recognit. 3360–3367.

Ward, B.D., 1999. 3d Intracranial: Automatic Segmentation of Intracranial Region.

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

Zhuang, A.H., Valentino, D.J., Toga, A.W., 2006. Skull-stripping magnetic resonance brain images using a model-based level set. NeuroImage 32, 79–92.