•

Local Label Learning (LLL) for Subcortical Structure Segmentation: Application to Hippocampus Segmentation

Yongfu Hao,¹ Tianyao Wang,² Xinqing Zhang,³ Yunyun Duan,⁴ Chunshui Yu,⁵ Tianzi Jiang,¹ and Yong Fan,¹* for the Alzheimer's Disease Neuroimaging Initiative

 ¹Brainnetome Center, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
 ²Department of Radiology, Shanghai East Hospital, Shanghai, China
 ³Department of Neurology, Xuanwu Hospital of Capital Medical University, Beijing, China
 ⁴Department of Radiology, Xuanwu Hospital of Capital Medical University, Beijing, China
 ⁵Department of Radiology, Tianjin Medical University General Hospital, Tianjin, China

Abstract: Automatic and reliable segmentation of subcortical structures is an important but difficult task in quantitative brain image analysis. Multi-atlas based segmentation methods have attracted great interest due to their promising performance. Under the multi-atlas based segmentation framework, using deformation fields generated for registering atlas images onto a target image to be segmented, labels of the atlases are first propagated to the target image space and then fused to get the target image segmentation based on a label fusion strategy. While many label fusion strategies have been developed, most of these methods adopt predefined weighting models that are not necessarily optimal. In this study, we propose a novel local label learning strategy to estimate the target image's segmentation label using statistical machine learning techniques. In particular, we use a L1-regularized support

Alzheimer Immunotherapy Research & Development, LLC; Johnson & Johnson Pharmaceutical Research & Development LLC; Medpace, Inc.; Merck & Co., Inc; Meso Scale Diagnostics, LLC; Novartis Pharmaceuticals Corporation; Pfizer Inc; Servier; Synarc Inc.; and Takeda Pharmaceutical Company; NIH; Contract grant number: P30 AG010129; K01 AG030514; Contract grant sponsor: Dana Foundation; Contract grant sponsor: The Canadian Institutes of Health Research, Foundation for the National Institutes of Health (available at: www.fnih.org), and Contract grant sponsor: Northern California Institute for Research and Education.

Received for publication 15 April 2013; Accepted 17 June 2013.

DOI: 10.1002/hbm.22359

Published online 23 October 2013 in Wiley Online Library (wileyonlinelibrary.com).

© 2013 Wiley Periodicals, Inc.

This article was published online on 23 October 2014. An error was subsequently identified. This notice is included in the online and print versions to indicate that both have been corrected 28 February 2014.

Contract grant sponsor: National Basic Research Program of China (973 Program); Contract grant number: 2011CB707801; Contract grant sponsor: Hundred Talents Program of the Chinese Academy of Sciences; Contract grant sponsor: National Science Foundation of China; Contract grant number: 30970770; 91132707; and 60831004; Contract grant sponsor: Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health); Contract grant number: U01 AG024904; Contract grant sponsor: National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd; AstraZeneca; Bayer HealthCare; BioClinica, Inc; Biogen Idec Inc; Bristol-Myers Squibb Company; Eisai Inc; Elan Pharmaceuticals Inc; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc; GE Healthcare; Innogenetics, N.V; Janssen

^{*}Correspondence to: Yong Fan, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China. E-mail: yong.fan@ieee.org, yfan@nlpr.ia.ac.cn

vector machine (SVM) with a k nearest neighbor (kNN) based training sample selection strategy to learn a classifier for each of the target image voxel from its neighboring voxels in the atlases based on both image intensity and texture features. Our method has produced segmentation results consistently better than state-of-the-art label fusion methods in validation experiments on hippocampal segmentation of over 100 MR images obtained from publicly available and in-house datasets. Volumetric analysis has also demonstrated the capability of our method in detecting hippocampal volume changes due to Alzheimer's disease. *Hum Brain Mapp* 35:2674–2697, 2014. © 2013 Wiley Periodicals, Inc.

Key words: multi-atlas based segmentation; local label learning; hippocampal segmentation; SVM

INTRODUCTION

Subcortical structure segmentation from magnetic resonance (MR) brain images is of great importance in a variety of neuroimaging studies, such as the brain anatomy and function [Mazziotta et al., 1991], the brain development, and brain disorders [Ostby et al., 2009; Sowell et al., 2002]. However, it is a challenging task to achieve automatic segmentation of subcortical structures from MR images due to a high degree of overlap between their intensity distributions and blurred boundaries between subcortical regions and background [Fischl et al., 2002; Tu et al., 2008].

Among existing medical image segmentation techniques, atlas-based methods have attracted great attention partially because spatial positions of anatomical structures are relatively stable across subjects. Given a target image to be segmented, the atlas-based methods spatially register one atlas image to the target image so that its associated atlas label is propagated with the obtained deformation field to the target image space [Bajcsy et al., 1983; Collins et al., 1995; Gee et al., 1993; Iosifescu et al., 1997]. The performance of such methods hinges on image registration accuracy and anatomical differences between the target and atlas images.

It has been proposed to select the most similar atlas to the target image based on either image similarity [Aljabar et al., 2009] or demographic information [Hajnal et al., 2007] in the atlas based image segmentation with multiple available atlases for alleviating the impact of anatomical variability. Although such strategies have been shown to be able to improve image segmentation performance [Avants et al., 2010; Wu et al., 2007], they might not work effectively if the target image is much different from all available atlases. Multiple atlases can also be used to generate a probability map as a prior information in statistical image segmentation algorithms [Ashburner and Friston, 2005; Collins et al., 1999; Fischl et al., 2004; Han and Fischl, 2007; Leventon et al., 2000; Marroquin et al., 2003; Pohl et al., 2006a; Twining et al., 2005; Yeo et al., 20081.

Recent studies have demonstrated that multi-atlas based segmentation methods can achieve robust performance by

fusing propagated labels of multiple atlases in the target image space [Artaechevarria et al., 2009; Heckemann et al., 2006; Khan et al., 2011; Rohlfing and Maurer, 2007; Rohlfing et al., 2004b; Sdika, 2010; Warfield et al., 2004]. A multi-atlas based segmentation algorithm typically consists of two steps: (1) registering each atlas image to the target image so that the atlas label is propagated to the target image space, and (2) fusing all the propagated atlas labels to generate a segmentation result of the target image. For achieving improved multi-atlas based image segmentation performance, besides optimizing the image registration [Heckemann et al., 2010; Jia et al., 2012; Khan et al., 2008, 2009], many methods have been proposed to improve the label fusion. Among the existing label fusion methods, probably the most simple and intuitive one is majority voting [Heckemann et al., 2006; Rohlfing et al., 2004a; Rohlfing and Maurer, 2007]. Improved label fusion strategies have been proposed to take into account local or global similarity between the target and atlas images in a weighted linear combination framework, such as image similarity based atlas weighting [Artaechevarria et al., 2008, 2009; Isgum et al., 2009; Sabuncu et al., 2010], segmentation performance based atlas weighting [Asman and Landman, 2011; Rohlfing et al., 2004b; Warfield et al., 2004], and regression based atlas weighting [Khan et al., 2011; Wang et al., 2011b]. Particularly, it has been demonstrated that linear combination models with their weights derived from patch-based image similarity measures are robust in several image segmentation studies [Coupe et al., 2011; Rousseau et al., 2011]. To relieve the adverse effects of atlases that are much different from the target image and reduce the computation cost, an atlas selection procedure, a special case of atlas weighting with binary weights, can also be adopted in multi-atlas based segmentation algorithms [Aljabar et al., 2009; Langerak et al., 2010; Leung et al., 2010; Lotjonen et al., 2010; van Rikxoort et al., 2010].

Along with the multi-atlas based methods for subcortical structure segmentation, supervised learning based image segmentation methods have been proposed to build classifiers based on the information of multiple atlases. Such supervised learning based segmentation methods first extract image features with information often richer than intensity information alone, and then construct a classification model based on the image features using supervised learning algorithms, such as SVMs [Morra et al., 2010; Powell et al., 2008], boosting [Morra et al., 2010], and artificial neural networks [Magnotta et al., 1999; Pierson et al., 2002; Powell et al., 2008; Spinks et al., 2002]. Since subcortical structures often exhibit complex appearance patterns, such as similar appearance of different subcortical regions and distinct appearance in different parts of the same subcortical region, one single classification model cannot successfully capture subcortical structures' complex appearance patterns in different locations, despite efforts to address this problem by combining shape or context information [Morra et al., 2009b; Tu et al., 2008; Tu and Toga, 2007; Wang et al., 2011a].

In this article, a novel local label learning (LLL) framework is proposed for image segmentation based on multiple atlases spatially registered to the target image using a pairwise non-linear image registration algorithm. In particular, for each of the target image voxels, a classification model for segmentation is learned from voxels of atlas images within a spatial neighborhood of the voxel considered. The classification model is learned using a hybrid of SVM and kNN (k nearest neighbor classification) [Zhang et al., 2006], which builds a SVM classifier for the voxel considered based on its k nearest positive and negative training samples. For reducing the computational cost, a probabilistic voting of the labels of registered atlases is adopted to identify image voxels with 100% certainty, and the classification is applied to those image voxels with uncertain probabilistic label voting. Our algorithm for segmenting hippocampus has been validated on 117 MR images of different scanning field strengths (1.5 T and 3.0 T) and different diagnostic groups of Alzheimer's disease (AD) as well as MR images of different scanning field strengths (1.5 T and 3.0 T) of 50 epilepsy patients and normal subjects. The experimental results indicated that our method could achieve competitive performance compared with state-of-the-art multi-atlas based segmentation methods. Preliminary results of this study have been reported in [Hao et al., 2012]. The software of our method will be made publicly available at http:// www.nitrc.org, and imaging data along with the manual segmentation results will be distributed with the software.

MATERIALS AND METHODS

Subjects and Imaging

Some of the image data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a \$60 million, 5-year public–private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

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

From the ADNI database, 30 subjects were randomly selected and they were equally distributed in three diagnostic groups, i.e., 10 patients with Alzheimer's disease (AD), 10 subjects with mild cognitive impairment (MCI), and 10 normal control people (NC). For each of them, both 1.5 T and 3.0 T T1-weighted MR images were downloaded. The dataset of 1.5 T images is referred to as Dataset A, and the dataset of 3.0 T images is referred to as Dataset B. For all the images, corrections including Grad-Warp [Jovicich et al., 2006], B1-correction [Jack et al., 2008], "N3" bias field correction [Sled et al., 1998] and geometrical scaling [Jack et al., 2008], were performed by ADNI.

An in-house dataset, referred to as Dataset C, consisting of Sagittal T1-weighted MR images of 57 subjects (20 NC, 15 MCI, and 22 AD), was acquired using a 3.0 T Siemens scanner with a magnetization prepared rapid gradient echo (MP-RAGE) sequence (TR/TE = 2,000/2.6 ms; FA = 9°; slice thickness = 1 mm, no gap). "N3" bias field correction [Sled et al., 1998] was applied to the images for reducing intensity inhomogeneity. These subjects' clinical scores and demographic information are shown in Table I.

Manual segmentation of these images was performed by two trained experts, one for Dataset A and Dataset B, and the other for Dataset C. The manual delineation was performed on coronal slices using ITK-snap [Mazziotta et al., 1991] following two protocols, one for the delineation of hippocampal head and body [Ostby et al., 2009] and the other for the delineation of the hippocampal tail [Sowell et al., 2002]. Ten images were randomly selected for

	AI	DNI (dataset A and	l B)	Ι	n-house (dataset (2)
	NC	MCI	AD	NC	MCI	AD
Subject Size	10	10	10	20	15	22
Age (years)	73.2(6.3)	73.5 (9.6)	73.1(7.0)	65.4(8.3)	71.5(9.2)	65.1(7.3)
Males/Females	5/5	5/5	5/5	7/13	7/8	12/10
MMSE	29.3(1.3)	26.1(1.8)	22.2(1.9)	28.3(1.5)	22.3(3.8)	9.1(6.5)
Manufacturer (SIEMENS/GE)	6/4	7/3	6/4	20/0	15/0	22/0

TABLE I. Demographic data and clinical scores of the subjects

Dataset A and Dataset B are the ADNI subjects' 1.5 T and 3.0 T scans, respectively.

accessing the intra-rater and inter-rater reliability. In terms of Dice index (see definition in following sections), the intra-rater reliability was 0.91 for both experts and the inter-rater variability was 0.89, similar to those reported in [Morra et al., 2009b; Wang et al., 2011a].

Another set of MR images of epilepsy patients and normal subjects were obtained from a publicly available dataset (http://www.radiologyresearch.org/HippocampusSeg mentationDatabase/) [Jafari-Khouzani et al., 2011]. This dataset consisted of T1 weighted MR images of 50 subjects, including 40 epileptic (13 males, 27 females; age range 15–64) and 10 nonepileptic subjects (five males, five females; age range 19–54). Images were acquired using two different MR imaging systems with different field strengths (30 1.5 T images with pixel size $0.78 \times 0.78 \times 2.00$ mm³, and 20 3.0 T images with pixel size $0.39 \times 0.39 \times 2.00$ mm³). Twenty-five images were selected for training and were provided with hippocampal labels. The other 25 images were provided without labels for testing algorithms [Jafari-Khouzani et al., 2011].

Multi-Atlas Based Image Segmentation

Given a target image *I* to be segmented and *N* atlases $\tilde{A}_i = (\tilde{I}_i, \tilde{L}_i), i=1, \ldots, N$, where \tilde{I}_i is an image and \tilde{L}_i is its associated segmentation label with value +1 indicating foreground and -1 indicating background, a multi-atlas based image segmentation algorithm first spatially registers the atlas images to the target image, then propagates the atlas labels to the target image space using the obtained deformation fields, and finally fuses the propagated atlas labels to generate a segmentation result using a specific label fusion strategy. For simplicity, we use $A_i = (I_i, L_i)$ to denote an atlas that has been spatially registered to the target image. The label of a target image voxel *x* can be computed as

$$\widehat{L}(x) = \underset{i \in \{-1,+1\}}{\arg \max} \sum_{i=1}^{N} w_i(x) p(L(x) = l | L_i),$$
(1)

where $w_i(x)$ is a weight assigned to the atlas label L_i at position x and l indicates the possible labels -1 or +1,

 $p(L(x)=l|L_i)$ is the probability that *L* belongs to label *l* at *x* given an atlas label L_i . One of the frequently used probability estimations of $p(L(x)=l|L_i)$, can be formulated as

$$p(L(x)=l|L_i) = \begin{cases} 1, & \text{if } L_i(x)=l \\ 0, & \text{otherwise} . \end{cases}$$
(2)

Many label fusion methods have been proposed and the major difference among them lies in how to define the weight $w_i(x)$. Depending on the weighting scheme used, label fusion methods typically fall into one of two categories: weights computed based solely on atlas labels and weights computed based on both atlas labels and image appearance.

The methods in the first category utilize the atlas labels solely to estimate the final segmentation. The simplest method is majority voting [Heckemann et al., 2006; Rohlfing and Maurer, 2007], which assumes that each atlas contributes equally to the image segmentation. The majority voting typically results in a binary label for classification [Kittler et al., 1998]. In this article, we refer to a summation of the individual labels, encoded by a probabilistic value, as probabilistic voting, which can be thresholded with a value, e.g., 0.5, to get a binary label. For image segmentation, with *N* atlases, the weights used in the probabilistic voting can be formulated as

$$w_i(x) = 1/N. \tag{3}$$

Shape-based averaging can also be viewed as a special case of probabilistic voting in that the atlas labels are transformed into Euclidean distance maps [Rohlfing and Maurer, 2007], which is similar to majority voting with a continuous probability estimation [Pohl et al., 2006b].

Another representative method in the first category is simultaneous truth and performance level estimation (STA-PLE) which was proposed to fuse segmentation results of multiple raters by simultaneously estimating their performance so that different weights can be assigned to raters according to their performance in label fusion [Rohlfing et al., 2004b; Warfield et al., 2004]. Although STAPLE has achieved great success in characterizing the performance of the raters, it sometimes yields worse results than majority voting for the multi-atlas based image segmentation [Artaechevarria et al., 2009; Khan et al., 2011; Langerak et al., 2010]. The performance of STAPLE might be improved by making it spatially adaptive [Asman and Landman, 2011], introducing more variables to characterize the label performance [Asman and Landman, 2011; Gering et al., 2001], or combining it with atlas selection strategy [Langerak et al., 2010]. However, all of these methods ignore the atlas image's appearance information that might be useful for achieving robust image segmentation.

The methods in the second category utilize both image appearance and label information of the atlases to fuse multiple labels. Specifically, a similarity measure of image appearance between the target and atlas images is typically used to determine weights for different atlases. The image appearance similarity can be measured globally for the whole image or locally for each voxel separately [Artaechevarria et al., 2009; Sabuncu et al., 2010; Wang et al., 2011b]. The global similarity based weighting strategy can be seen as a generalized version of majority voting or STAPLE, where the contribution of each atlas label to the final segmentation is proportional to its global image appearance similarity to the target image [Artaechevarria et al., 2009]. The local similarity based weighting methods adaptively assign a weight to each atlas voxel separately. A Gaussian weighting model (LWGU) with summed square distance (SSD) has been proposed in [Sabuncu et al., 2010; Wang et al., 2011b], and the local weight can be computed for each atlas voxel by

$$w_{i}(x) = \exp\left(-\sum_{y \in N(x)} \frac{||I_{i}(y) - I(y)||_{2}^{2}}{\sigma_{x}}\right),$$
 (4)

where N(x) defines a spatial neighborhood of voxel x, $|| \cdot ||_2$ is the Euclidean distance between intensities of $I_i(y)$ and I(y), and σ_x is a parameter of the weighting model. Similarly, a spatially adaptive weight can be computed using the inverse weighting model (LWINV) as defined in [Artaechevarria et al., 2009; Wang et al., 2011b] by

$$w_{i}(x) = \left[\sum_{y \in N(x)} ||_{Ii}(y) - I(y)||_{2}^{2}\right]^{p},$$
(5)

where p is a parameter of the weighting model with a negative value.

The label fusion can also be achieved based on a patch based weighting scheme, referred to as nonlocal patch based fusion (NLP), as defined in [Coupe et al., 2011; Rousseau et al., 2011] by

$$\hat{L}(x) = \frac{\sum_{S=1}^{N} \sum_{j \in V} w(x, x_{s,j}) \cdot L(x_{s,j})}{\sum_{S=1}^{N} \sum_{j \in V} w(x, x_{s,j})}$$
(6)

where $L(x_{s,j})$ is the label of voxel $x_{s,j}$ at location j in atlas s, V is a search volume, and $w(x, x_{s,j})$ is the weight assigned to $L(x_{s,j})$ based on the similarity between the patches surrounding x and $x_{s,j}$. In particular, the weight $w(x, x_{s,j})$ can be computed as [Coupe et al., 2011]:

$$w(x, x_{s,j}) = \begin{cases} e^{-\frac{||p(x) - p(x_{s,j})||_2^2}{\sigma_x}} & \text{if } ss > th \\ 0 & \text{else} \end{cases}$$

where $P(\cdot)$ is a cubic patch centered at the voxel considered, and $|| \cdot ||_2$ is L2 norm computed between each intensity of the elements of the patches P(x) and $P(x_{s,j})$, *ss* is a structure similarity measure between the two patches [Coupe et al., 2011], and *th* is a threshold. A similar strategy has been used for incorporating image intensity information in STAPLE [Asman and Landman, 2012].

Recently, a generative model employing a nonparametric estimator has been proposed for estimating the posterior label probability of the target image [Sabuncu et al., 2010]. Within this framework, several existing label fusion methods, such as majority voting, global similarity based weighing and LWGU, can be treated as one of its special cases.

Most of the existing image similarity based local weighting methods explicitly specify a weighting scheme that is not necessarily the best fit for the label fusion problem. To automate the determination of atlas weights for label fusion, a regularized linear model of propagated segmentation labels weighted by image appearance difference between the target and atlas images has been built by least square fitting [Khan et al., 2011]. Correlations between results produced by different atlases have also been taken into account for label fusion [Wang et al., 2011b].

Local Label Learning Method (LLL)

As schematically illustrated in Figure 1, the local label learning (LLL), for directly learning a classifier for each voxel of the target image to be segmented, consists of candidate training set construction, feature extraction, and local SVM classification.

Candidate training set construction

To learn an image segmentation classifier for each voxel of the target image, a set of voxel-wise training samples is identified from the registered atlases. Since the image registration of atlases cannot achieve perfect alignment of all image voxels across images, it is not appropriate to directly take the corresponding voxel of the voxel considered in each atlas as a training sample. To achieve better correspondence between voxels of the target and atlas images, a local search strategy can be used to find the best match in each atlas for the target image voxels [Wang et al., 2011b]. However, the local searching for the best



Figure I.

The framework of local label learning (LLL) method, consisting of steps: (1) candidate training set construction, (2) feature extraction, and (3) local SVM classification. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

match is computationally expensive if not impossible. Such a strategy also limits the number of training samples obtained no more than the number of atlases, which may require a special treatment for learning algorithms in studies with a limited number of atlases [Wang et al., 2011b].

Instead of obtaining only one sample from each atlas, we adopt a local patch based method [Coupe et al., 2011; Rousseau et al., 2011]. Given one voxel x of the target image, as illustrated in Figure 1, voxels in its neighborhood N(x) of all atlases are used as training samples. Particularly, we take a $(2r+1)\times(2r+1)\times(2r+1)$ cube-shaped neighborhood and get $N \cdot (2r+1)^3$ candidate training samples $\left\{ \left(\vec{f}_{i,j}, l_{ij}\right) | i=1, \ldots, N, j \in N(x) \right\}$ from N atlases, where r is the neighborhood radius, $\vec{f}_{i,j}$ is a feature vector extracted from voxel j of the ith atlas by the feature extraction method to be described next, and each candidate training sample's segmentation label is $l_{i,j} \in \{+1, -1\}$, same to its atlas label. When r=0, only the corresponding voxel in each atlas is used as a training sample.

The candidate training samples from a local patch have different degrees of similarity to the target voxel to be labeled. A hybrid of SVM and *k*NN can identify balanced positive and negative training samples, the most similar to the target image voxel considered, and learn an effective and efficient SVM classifier based on the balanced training samples. It is worth noting that unlike the affine image registration used in [Coupe et al., 2011; Rousseau et al., 2011], we adopt a non-rigid image registration algorithm for registering the target and atlas images so that a smaller neighborhood can be used in our algorithm, which makes the trained classifier more resistant to noise/outliers.

Feature extraction

As demonstrated in several subcortical segmentation studies [Fischl et al., 2002; Tu et al., 2008], image intensity information solely is not good enough for distinguishing different subcortical structures since most subcortical structures share similar intensity patterns in MR images. To address such a problem, in the learning based segmentation methods, more discriminative features are often extracted from MR images [Morra et al., 2010; Powell et al., 2008; Tu et al., 2008]. However, such a strategy has not been widely employed for label fusion in atlas-based segmentation studies.

In our method, a set of features is extracted for capturing texture information of subcortical structures [Toriwaki and Yoshida, 2009]. The features extracted for each voxel include intensities in its neighborhood of size $(7 \times 7 \times 7)$, and outputs of the first order difference filters (FODs), the second order difference filters (SODs), 3D Hyperplane filters, 3D Sobel filters, Laplacian filters, and Range difference filters [Toriwaki and Yoshida, 2009]. All of them are concatenated to form a feature vector \vec{f} with 379 elements.

Given an image *I*, FODs and SODs, capable of detecting intensity change along a line segment, can be computed as $I(x+r\cos\theta\sin\phi, y+r\sin\theta\sin\phi, z+r\cos\phi) - I(x-r\cos\theta\sin\phi, y-r\sin\theta\sin\phi, z-r\cos\phi)$ and $I(x+r\cos\theta\sin\phi, y+r\sin\theta\sin\phi, z+r\cos\phi) + I(x-r\cos\theta\sin\phi, y-r\sin\theta\sin\phi, z-r\cos\phi) - 2I(x, y, z)$, respectively, where I(x, y, z) is *I*'s intensity at voxel (x, y, z), θ and θ are two rotation angles, *r* is the distance from (x, y, z) to the voxels considered. In this study, $r=1, \theta \in \{0, \pi/4, \pi/2, 3\pi/4\}$, and $\phi \in \{0, \pi/4, \pi/2\}$ were used for the first and second order difference filters.

Three-dimensional Hyperplane filters and 3D Sobel filters are extensions of the first order difference filters and can be formulated as $F_{ijk} = P * (I_{i+1} - I_{i-1})$, where *P* is a 3×3 $\begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$

kernel operator (for Hyperplane filters $P = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$,

while for 3D Sobel filters
$$P = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 3 & 2 \\ 1 & 2 & 1 \end{bmatrix}$$
), * denotes convo-

lution operation, I_{i-1} and I_{i+1} denote two planes along x axes of image I at position i-1 and i+1. It is worth noting that the above two filters are directional. Filters along directions of y and z can be similarly implemented.

Laplacian filters are isotropic and can be treated as extensions of the second order difference filters. They can be formulated as $F_{(x,y,z)} = \sum_{l,m,n \in N(x,y,z)/(x,y,z)} (I(l,m,n)-I(x,y,z))$ where N(x,y,z)/(x,y,z) denotes voxels in the neighborhood of voxel (x,y,z) excluding itself. Depending on the number of voxels used, Laplacian filter is also referred to as *p*-neighbor Laplacian.In this study, three filters, namely 6–neighbor Laplacian, 18–neighbor Laplacian and 26-neighbor Laplacian, implemented in the $3 \times 3 \times 3$ neighborhood were used for feature extraction.

The range difference filter computes the difference between maximal and minimal values in a given neighborhood of each voxel and can be formulated as $F_{(x,y,z)}$ =max $(I_{N(x,y,z)})$ -min $(I_{N(x,y,z)})$, where N(x,y,z) denotes voxels in a given neighborhood of (x, y, z), max (·) and min (·) get



Figure 2.

Feature extraction for a randomly selected image by applying filters with different parameters. The displayed filtering outputs are scaled to have the same intensity range as the intensity indicator.

the maximal value and the minimal value of the input, respectively. Three neighborhood sizes, including 7-neighbor, 19-neighbor, and 27-neighbor, extracted in the $3\times3\times3$ neighborhood were used in this study for extracting features. Some filtering outputs of a randomly selected image are shown in Figure 2.

In this study, to account for difference of intensity distributions across atlases and the target image, before feature extraction for voxel v, we normalized the intensities of voxels in a cube patch centered atv with a neighborhood size of $7 \times 7 \times 7$. Particularly, mean value and standard deviation of intensities of all voxels in the cube patch were first computed, and then the intensity of each voxel in this cube patch was subtracted by the mean value and divided by the standard deviation.

Local SVM classification

Once the image features are extracted for both the target image voxel v and its corresponding candidate training samples, including n_p positive samples and n_n negative samples, a *k*NN strategy based SVM classification algorithm is adopted to build a classifier [Zhang et al., 2006].

In particular, approximately *k* nearest neighboring samples of the target image voxel are identified from the candidate training samples based on Euclidean distances between their feature vectors and the target image voxel's feature vector, including $k_{pos} = \min \{floor(\frac{k}{2}), n_p\}$ positive nearest neighboring samples and $k_{n \in g} = \min \{floor(\frac{k}{2}), n_n\}$ negative nearest neighboring samples, constituting a balanced training set, as illustrated by samples in the dashed black circle in Figure 1.

A SVM classifier is built on the identified training set and then used to classify the target voxelv. As the features might contain redundant information, we adopt an L1regularized SVM method for obtaining a sparse model [Yuan et al., 2010]. The L1-regularized SVM classifier can be obtained by solving an optimization problem:

$$\min \vec{w} ||\vec{w}||_1 + C \sum_{i,j} \left(\max \left(0, 1 - l_{i,j} \vec{w}^T \vec{f}_{i,j} \right) \right)^2, \tag{7}$$

where $|| \cdot ||_1$ donates L1 norm.

The L1-regularized SVM optimization problem can be solved by a coordinate descent method [Yuan et al., 2010]. In particular, our algorithm implementation used a publicly available software package LIBLINEAR (www.csie.ntu.e-du.tw/~cjlin/liblinear/) [Fan et al., 2008]. The L1-regularized SVM often produces a sparse solution of \vec{w} , and nonzero elements of \vec{w} are informative features selected. Once we get \vec{w} , the label of target image voxel x can be estimated as

$$\hat{L}(x) = \operatorname{sgn}\left(\vec{w}^T \vec{f}\right).$$
(8)

The LLL algorithm is summarized as

LLL ALGORITHM

Inputs: One target image *I* to be segmented, *N* atlases $A_i = (I_i, L_i), i=1, ..., N$ that have been spatially regis-

tered to the target image space, where I_i is the *i*th atlas image and L_i is its associated segmentation label.

Output: Label map *L* of the target image *I*. *Begin:*

For each voxel *x* in the target image

- Obtain $N \cdot (2r+1)^3$ candidate training samples and compute their image feature vectors.
- Find *K*_{pos} positive training samples and *K*_{neg} negative training samples respectively from the candidate training set based on Euclidean distances between feature vectors of the testing and training samples.
- Training a classifier based on the selected $K_{\text{pos}} + K_{\text{neg}}$ training samples using L1-regularized SVM with a linear kernel.
- Estimate label L(x) of the target image voxel x by applying the SVM classifier to its feature vector.

End for End

Segmentation of Hippocampus

All MR images were registered to the MNI152 space using affine registration, and these aligned images were resampled to have a voxel size of $1 \times 1 \times 1$ mm³. For each of the left and right hippocampi, a bounding box was generated to cover the whole hippocampus in the MNI152 space following the procedure described in [Morra et al., 2009b]. In particular, all the atlases were scanned to find the minimum and maximum *x*, *y*, *z* positions of the hippocampus and the size of seven voxels was added in each direction to cover the hippocampus of unseen testing images.

Atlas Selection and Image Registration

For segmenting a target image, an atlas selection strategy was adopted to select the most similar atlases. The similarity between the target image and an atlas image was evaluated based on normalized mutual information (NMI) of image intensities within the bounding box [Aljabar et al., 2009]. In this study, we chose the top 20 most similar atlases [Morra et al., 2009b]. All the selected atlas images were independently registered to the target image using a nonlinear, cross-correlation-driven image registration algorithm, namely ANTs (version 1.9x) [Avants et al., 2008], with the command: ANTS 3 -m CC [target.nii, source.nii, 1, 2] –i 100x100x10 –o output.nii -t SyN[0.25] –r Gauss[3,0].

Initial Segmentation With Probabilistic Voting

To reduce the computation cost, the probabilistic voting strategy was adopted to get an initial segmentation result of the target image. For each voxel of the target image, we got the probability value of the voxel belonging to the hippocampus. The segmentation result of voxels with 100% certainty (probability value of 1 or 0) was directly taken as the final segmentation result. Then the local label learning focused on voxels with probability values greater than 0 and smaller than 1.

Parameter Tuning

The parameters of our algorithm were determined empirically based on Dataset A, including the neighborhood radius *r* for obtaining the candidate training samples and the parameter *k* for *k*NN. A leave-one-out validation was adopted to tune the parameters for achieving the optimal segmentation performance. Particularly, the searching range of *r* was {0, 1, 2, 3}, *k* was selected from {200, 300, 400, 500}. In particular, when r=0, all available training samples were used in the SVM classifier training. We used the default value for parameter C of the L1-regulzaried SVM as recommended in [Fan et al., 2008]. We also performed the segmentation by replacing the SVM classifier with a *k*NN classifier. The performance of the a *k*NN classifier based segmentation was evaluated with *k* selected from {1, 5, 10, 20, 50, 100, 150, 200, 250, 300, 400, 500}.

Feature Selection Analysis

To investigate which features are informative for the classification, we identified features most frequently selected by SVM classifiers for the segmentation of right hippocampus for images of Dataset A. As classifiers at different locations of the hippocampal structure may select different features, we focused on classifiers for voxels located at the boundary of the hippocampus. In particular, for each image of Dataset A, we first identified the boundary of its hippocampal structure, then found the features selected by the classifiers, i.e., those corresponding to non-zero elements of \vec{w} , for different boundary voxels, and finally we obtained the frequency of features selected by different classifiers. The mean of frequencies of features selected for the segmentation of images of Dataset A was also obtained.

Comparison With State-of-the-Art Label Fusion Methods

The proposed algorithm was compared with state of the art label fusion algorithms, including majority voting, STAPLE, LWGU [Sabuncu et al., 2010], LWINV [Wang et al., 2011b], and NLP [Coupe et al., 2011; Rousseau et al., 2011]. The parameters of LWGU, LWINV, and NLP were tuned based on Dataset A using cross-validation. For LWGU, patch radius *r* and σ_x need to be determined. In particular, σ_x was adaptively set to $\min_{y \in N(x)} \{|I(y) - I_i(y)|^2 + \epsilon\}$, i=1,...,N, and ϵ was set to 1e-20 for numerical stability. The optimal value of *r* was 1, selected from {1, 2, 3}. For LWINV's parameters, including *p* and patch radius *r*,

the optimal value of p was -1, selected from $\{-3, -2, -1\}$, while the optimal value of r was 1, selected from $\{1, 2, 3\}$. For NLP, there are three parameters, including patch radius r, search volume V, and σ_x . The optimal value of r was 1, selected from $\{1, 2, 3\}$. The optimal value of V was $3\times3\times3$, selected from $\{3\times3\times3, 5\times5\times5, 7\times7\times7\}$, σ_x was adaptively set to min $_{y\in N(x)}\{|P(x)-P(x_{s,j})|_2^2+\epsilon\}$, and ϵ was set to 1e-20 for numerical stability. The NLP label fusion method was performed based on atlases nonlinearly registered to the target image space, instead of those registered with affine transformation [Coupe et al., 2011; Rousseau et al., 2011].

Validation of the Segmentation Performance Across Different Datasets

We evaluated our method across different datasets. In particular, we segmented each image of Dataset A (1.5 T SIE-MENS/GE scanners) with atlas images obtained from Dataset B (3.0 T SIEMENS/GE scanners) excluding that from the same subject of the image of Dataset A under consideration, and segmented each image of Dataset B with atlas images obtained from Dataset A excluding that from the same subject of the image of Dataset B under consideration. It is worth noting that images of Dataset A and Dataset B were acquired with different scanners and from subjects of different diagnostic groups, including AD, MCI, and NC. We also segmented each image of Dataset C (3.0 T SIEMENS scanner) with atlas images obtained from Dataset A (1.5 T SIEMENS/ GE scanners) or Dataset B (3.0 T SIEMENS/GE scanners).

Validation of the Segmentation Performance Based on Public Available Data

Besides MR images from studies of Alzheimer's disease, we also validated our method based on a publicly available dataset of nonepileptic subjects and epilepsy patients [Jafari-Khouzani et al., 2011]. Using the training data of 25 subjects as atlases, we obtained segmentation results of the testing data using our method, LWGU, LWINV, and NLP with the same parameters used in the previous experiments. As these images have anisotropic spatial resolution, we first interpolated the voxel size to $1 \times 1 \times 1$ mm³ before segmentation. Then, after the segmentation, we interpolated the image back to its original resolution for evaluating the segmentation performance. The performance of these methods was evaluated by the dataset provider based on 11 metrics and compared with two algorithms: Brain Parser [Tu et al., 2008] and a multi-atlas based segmentation method [Aljabar et al., 2007].

Evaluation Metrics

The segmentation performance of each method was evaluated using leave-one-out cross-validation. We adopted the metrics used in [Jafari-Khouzani et al., 2011] except "Specificity" due to the reason that it varies with the size of background. Given manual segmentation label E and the automated segmentation result F, these metrics can be calculated as:

- 1. Dice Index = $2 \frac{V(E \cap F)}{V(E) + V(F)}$
- 2. Jaccard Index = $\frac{V(E \cap F)}{V(E \cup F)}$
- 3. Precision = $\frac{V(E \cap F)}{V(F)}$,
- 4. Recall = $\frac{V(E \cap F)}{V(E)}$
- 5. Relative volume Difference : $RVD = \frac{V(E) V(F)}{V(F)}$
- 6. Mean Distance : MD = mean $_{e \in BE}(\min_{f \in BF} d(e, f))$,
- 7. Hausdorff Distance: HD = max (H(E,F), H(F,E)), where $H(E, F) = \max_{e \in BE} (\min_{f \in BF} d(e, f))$,
- 8. Hausdorff 95 Distance (HD95) is similar to Hausdorff Distance, except that 5% data points with the largest distance are removed before the calculation,
- 9. Average Symmetric Surface Distance: ASSD = $(\text{mean}_{e \in BE}(\min_{f \in BF} d(e, f)) + \text{mean}_{f \in BF}(\min_{e \in BE} d(e, f)))$ /2,
- 10. Root Mean Square Distance: $RMSD = \sqrt{D_E^2 + D_F^2}$ $(\operatorname{card} \{BE\} + \operatorname{card} \{BF\}).$

In above metrics, BE denotes boundary voxels of the segmentation E, BF denotes boundary voxels of the segmentation F, $d(\cdot, \cdot)$ is the Euclidian distance between two $D_{\rm E}^2 = \sum_{e \in BE} \min_{f \in BF} d^2(e, f), \quad D_F^2 = \sum_{f \in BF} \min_{e \in BE} d^2$ points, (e,f), card $\{\cdot\}$ is the cardinality of a set, and V(X) is the volume of segmentation X.

The first four metrics have been widely used in image segmentation studies and they are related to each other. The relative volume difference metric can reveal the segmentation method's capability for detecting volume changes, although it does not directly measure the overlap between segmentation labels. The last five metrics, measuring surface distance between segmentation results, characterize the boundary difference.

Hippocampal Volumetry for Alzheimer's Disease

A volumetric analysis was carried out to test the capacity of each method for capturing the volume difference among NC, MCI, and AD, based on the available 3.0 T images in Dataset B and Dataset C. Based on the segmentation results, each subject's hippocampal volumes were computed and normalized by its total intracranial volume (TIV) estimated by VBM8 (http://dbm.neuro.uni-jena.de/vbm/download/). In particular, each subject's hippocampal volume was divided by its TIV and then multiplied by mean of TIVs of all subjects of Dataset B and Dataset C. We also performed two sample t tests, computed standardized effect sizes (Cohen's d), and estimated sample size of each group for different methods to detect a difference in total hippocampal volumes (left + right) between NC group and MCI group, as well as between NC group and AD group. For the sample size estimation, Type I error rate was set to 0.05, and power was set to 0.8 [Eng, 2003].

RESULTS

Initial Segmentation With Probabilistic Voting

The experiment results revealed that voxels labeled by the probabilistic voting [Eq. (3)] as either hippocampus



Figure 3.

The initial segmentation result of a randomly selected test image based on the probabilistic voting. The first row shows three slices of the test image with manual segmentation label. The probabilistic voting results are shown in the second row and the color bar indicates the probability of a voxel belonging to the hippocampus.

The voxels belonging to the hippocampus or not with 100% certainty (probability value is 1: red or 0: blue) are overlaid on the test image in the third row. First column: horizontal; second column: sagittal; third column: coronal. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

(+1) or background (-1) with 100% certainty were correctly segmented with a correct rate close to 100%. On average, only one voxel labeled with 100% certainty was misclassified in each image in Datasets A, B, and C. Figure 3 shows the initial segmentation result on one randomly selected test image. Therefore, our algorithm can focus on voxels labeled by the probabilistic voting with probability values greater than 0 and smaller than 1. In particular, about 98% of the background and 20% of the foreground voxels were labeled with 100% certainty for subjects of Dataset A within the bounding box of $44 \times 63 \times 64$ of one side of the hippocampus.

Parameter Tuning

Figure 4 shows average Dice index values of segmentation results of the right hippocampus with r varying from 0 to 3. For the SVM based segmentation, the segmentation performance improved first as the number of training samples K increased and reached its maximal when K was 400, then the performance degraded if K was greater than



Figure 4.

The average Dice index values of segmentation results of the right hippocampus for dataset A with different numbers of training samples and r varying from 0 to 3. Top: SVM classifier based segmentation results. Bottom: kNN classifier based segmentation, all available training samples (20 in total) were used without selection when r = 0. For the kNN classifier based segmentation, all available training samples were used. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

400. The optimal value of r was 1 and the optimal value for K was 400. The segmentation performance improvement with the increase of the number of training samples (<400) might be due to the relieved curse of dimensionality. However, the segmentation performance degraded if too many samples, e.g., 500 samples, were used in the SVM training, indicating that noisy training samples (located far away from the voxel considered) might be involved in the classifier training. The segmentation performance when r=0 was worse than when $r=\{1,2\}$, possibly due to that the limited training samples might not able to provide sufficiently discriminative information for building robust classifiers. The segmentation performance degraded as r varying from 1 to 3, indicating that irrelevant samples may have been used in the classifier training when larger searching radius *r* was used.

Similar to the SVM classifier based segmentation, the kNN classifier based segmentation performance improved and then degraded with the increase of k, the number of nearest neighbors used in the classification. The segmentation performance degraded as r varying from 1 to 3. However, the segmentation performance when r=0 was better than when $r=\{2,3\}$. The best Dice index was 0.877, achieved with 150 nearest neighbors and r=1. Overall, the segmenation performance of kNN classifiers was worse than that of SVM classifiers, suggesting that the sophisticated SVM classifiers should be used in the segmentation.

Feature Selection Analysis

The frequency of features used in the segmentation is shown in Figure 5. On average, about 7% features were selected for the segmentation. These results indicated that both intensity information and filtering outputs played important roles in the segmentation. The top ten most frequently selected features are summarized in Table II.

Comparison With State-of-the-Art Label Fusion Methods

Figures 6 to 8 show box plots of segmentation performance measures of majority voting, STAPLE, LWGU, LWINV, NLP, as well as LLL, based on Datasets A, B, and C, respectively. The mean values of the segmentation performance measures and P values of single-sided paired t tests for comparing LLL with others are reported in Tables III to V. All the results demonstrated that the proposed method performed consistently better than other label fusion methods.

Hippocampal volume differences between results of the manual segmentation and automatic segmentation are shown in Figure 9 and Table VI. Pearson correlation coefficients between hippocampal volumes estimated by the manual segmentation and by different methods are also reported in Table VI. These results demonstrated that our proposed method could better estimate the hippocampal volume, although small volume differences do not necessarily mean accurate segmentation results. For visual





Frequency of features used in the segmentation. (1) One randomly selected image with the hippocampus boundary voxels shown in different colors. (2) Features selected in the segmentation. Each row indexed by the color bar shown at the left corresponds to the boundary voxel in the same color shown in (1). The x-axis is the feature index (1–27: intensity features in the neighborhood of $3 \times 3 \times 3$, 1–125: intensity features in the neigh-

inspection, Figure 10 shows segmentation results of subjects randomly selected from each dataset.

The performance of STAPLE was relatively poor, as similarly reported in other atlas-based segmentation studies

 TABLE II. Top 10 most frequently selected features on average for the segmentation of dataset A

Rank	Feature type	Parameter			
1	SOD	$\theta = 3 \pi/4, \varphi = \pi/2$			
2	SOD	$\theta = \pi/2, \varphi = 0$			
3	SOD	$\theta = \pi/4, \varphi = \pi/2$			
4	Range filter	7-neighbor			
5	SOD	$\theta = \pi/2, \varphi = \pi/2$			
6	SOD	$\theta = 3\pi/4, \varphi = \pi/4$			
7	Laplacian filter	26-neighbor			
8	SOD	$\theta = 0, \varphi = \pi/2$			
9	SOD	$\theta = \pi/2, \varphi = \pi/4$			
10	Intensity	[4, 3, 5]			

borhood of $5 \times 5 \times 5$, 1:343: intensity features in the neighborhood of $7 \times 7 \times 7$, 344–379: filtering outputs). (3) Frequency of features selected for the segmentation of the image shown in (1). (4) Mean of frequencies of features selected for the segmentation of images of dataset A. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

[Artaechevarria et al., 2009; Khan et al., 2011; Langerak et al., 2010]. In our experiment, STAPLE had the worst performance evaluated by most of the performance metrics, except recall index and mean distance. The visual inspection and volume difference information revealed that STAPLE was prone to generating segmentation results larger than the manual segmentation results.

The overall performance of majority voting was better than STAPLE. Particularly, majority voting had a higher precision index and a lower recall index, indicating the hippocampal segmentation volume of this method was smaller than that of the manual segmentation compared with other methods, consistent with the results (negative RVD) reported in Table VI.

LWGU and LWINV had similar performance, better than both STAPLE and majority voting. This finding was consistent with results reported in [Artaechevarria et al., 2009; Sabuncu et al., 2010], and further justified the effectiveness of spatially adaptive image similarity measures for characterizing the image registration accuracy in multi-



Figure 6.

Box plots of the results for the dataset A. On each box, the central mark is the median, and edges of the box are the 25th and 75th percentiles. Whiskers extend from each end of the box to the adjacent values in the dataset and the extreme values within I interquartile range from the ends of the box. Outliers are data with values beyond the ends of the whiskers. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

atlas based image segmentation. NLP performed better than LWGU and LWINV, providing supportive evidence that better fusion performance can be obtained by relaxing the one-to-one correspondence constrains.

For all the datasets, the proposed method achieved the best overall performance. In particular, our method had the best value on most metrics, including Dice, Jaccard, MD, ASSD, and RMSD, confirmed by paired single-sided *t* test as summarized in Tables III, V. The Recall index of our method was just lower than STAPLE's, indicating that our method was able to generate segmentation results that better covered the foreground region of the manual segmentation. For most of the metrics, our method had fewer outliers compared with other methods, indicating that our method might be more robust than other methods. Our method produced segmentation volumes with the least difference from the manual segmentation results for most

datasets, except for the left hippocampus of Dataset A, as indicated by the results shown in Figure 9 and Table VI. It is worth noting that we presented measures of both RVD and absolute value of RVD. RVD itself contains directional information of the volume difference, and its absolute value reflects the absolute difference.

Validation of the Segmentation Performance Across Different Datasets

As shown in Table VII, for images obtained from ADNI, i.e., Dataset A and Dataset B, acquired with scanners of different makers and different field strengths, our method have achieved robust segmentation performance. However, the segmentation performance of our method for images of Dataset C with atlases obtained from Dataset A or Dataset B was worse than that obtained with atlases



Figure 7.

Box plots of the results for the dataset B. On each box, the central mark is the median, and edges of the box are the 25th and 75th percentiles. Whiskers extend from each end of the box to the adjacent values in the dataset and the extreme values within I interquartile range from the ends of the box. Outliers are data with values beyond the ends of the whiskers. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

obtained from Dataset C. Such a performance difference might be caused by different imaging protocols used for image acquisition and different ethnic groups of the subjects. The images of ADNI were acquired with a standard protocol at different imaging sites and geometric distortion of images were corrected. However, the images of Dataset C were acquired without correction of geometric distortion. The difference between with and without geometric distortion correction and the difference of ethnic groups might make the image registration difficult.

Validation on Public Available Data of Epilepsy Patients and Normal Subjects

Table VIII shows performance of the segmentation methods evaluated based on the publicly available dataset of epilepsy patients and normal subjects. Based on the published results [Jafari-Khouzani et al., 2011], our method performed much better than the supervised learning method [Tu et al., 2008] and the multi-atlas-based segmentation method [Aljabar et al., 2007]. Results of statistical tests demonstrated that our method was statistically better than three other methods (LWGU, LWINV, NLP) for 6 of 10 metrics. The degree of performance improvement achieved by our method over LWGU and LWINV was similar to those reported in [Artaechevarria et al., 2009; Khan et al., 2011]. However, the improvement was statistically significant as indicated by P values of paired-tests for comparing LLL with LWGU, LWINV, and NLP. These results also indicated that the segmentation methods performed better for images with isotropic voxels than those with highly anisotropic voxels, such as those in this public dataset (20 1.5 T images with voxel size $0.78 \times 0.78 \times 2.00$ mm³, five 3.0 T images with voxel size $0.39 \times 0.39 \times 2.00 \text{ mm}^3$).



Figure 8.

Box plots of the results for the dataset C. On each box, the central mark is the median, and edges of the box are the 25th and 75th percentiles. Whiskers extend from each end of the box to the adjacent values in the dataset and the extreme values within I interquartile range from the ends of the box. Outliers are data with values beyond the ends of the whiskers. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Hippocampal Volumetry

The distributions of hippocampal volumes estimated from the segmentation results of all 3.0 T images are shown in Figure 11, and group means and standard variances of NC, MCI, and AD are summarized in Table IX. These results demonstrated that the mean volumes of segmentation results produced by our method were the closest to those obtained by manual segmentation for NC group. The segmentation results produced by the automatic methods had smaller variance than those produced by the manual segmentation as indicated by the standard deviations of volumes estimated, suggesting that automatic methods could achieve segmentation results with better consistency. The *t* test results confirmed the finding that the hippocampal volume is a promising biomarker for Alzheimer's disease [Morra et al., 2009a; Schuff et al., 2009; Wolz et al., 2010b]. Our method could even detect a statistically significant difference between NC and MCI with a *P* value smaller than the manual segmentation's. Such findings were also confirmed by the effect sizes that characterize the standard mean volume difference between NC and its counterparts. As shown in Table IX, our method had the largest effect size and the smallest sample size of all automatic methods.

Effect of the Number of Atlases in Segmentation

In all above experiments, 20 atlases were used for segmenting the target image. To investigate how the number of atlases affects the segmentation performance, the segmentation performance associated with different numbers of atlases was evaluated for right hippocampus based on

	MV P value (L/R)	STAPLE P value (L/R)	LWGU P value (L/R)	LWINV P value (L/R)	NLP P value (L/R)	LLL
Dice (L/R)	0.860/0.868;	0.802/0.824;	0.872/0.881;	0.860/0.886;	0.882/0.888;	0.887/0.894
	3.4e-6/5.2e-10	5.2e-7/1.1e-6	1.7e-5/4.5e-6	1.9e-2/1.3e-3	3.3e-2/4.8e-2	
Jaccard (L/R)	0.755/0.767;	0.677/0.707;	0.775/0.789;	0.760/0.796;	0.789/0.799;	0.797/0.809
	1.6e-6/5.8e-10	1.1e-7/3.0e-7	1.4e-5/4.7e-6	1.3e-2/1.1e-3	3.7e-2/4.8e-2	
Precision (L/R)	0.894/0.900;	0.694/0.724;	0.864/0.898;	0.897/0.908;	0.898/0.899;	0.902/0.908
	2.7e-1/2.3e-1	1.2e-12/2.6e-12	4.2e-2/2.9e-2	3.5e-1/9.7e-1	3.2e-1/3.9e-2	
Recall (L/R)	0.834/0.843;	0.971/0.970;	0.863/0.876;	0.853/0.860;	0.870/0.880;	0.875/0.884
	1.3e-7/3.9e-8	1.0/1.0	2.1e-4/3.5e-3	1.0e-7/1.8e-7	8.5e-3/9.7e-2	
HD (mm) (L/R)	3.930/3.750;	4.100/3.990;	4.890/4.870;	3.700/3.730;	3.400/3.145;	3.610/2.901
	7.8e-1/1.6e-4	4.0e-1/1.2e-6	5.5e-2/1.3e-9	6.4e-1/7.0e-4	9.2e-1/3.9e-2	
HD95 (mm) (L/R)	1.820/1.780;	1.930/1.710;	2.220/2.300;	1.890/1.800;	1.187/1.097;	1.097/1.083
	3.1e-9/6.3e-11	3.7e-4/9.2e-4	2.4e-15/7.2e-16	1.5e-11/2.5e-10	4.2e-2/2.0e-1	
MD (mm) (L/R)	0.340/0.320;	0.470/0.440;	0.420/0.410;	0.390/0.370;	0.271/0.219;	0.280/0.264
	2.4e-3/2.9e-6	4.8e-11/1.0e-16	2.7e-6/5.1e-10	3.7e-5/3.9e-8	7.2e-3/2.4e-2	
ASSD (mm) (L/R)	0.330/0.290;	0.520/0.470;	0.410/0.390;	0.370/0.350;	0.300/0.250;	0.252/0.235
	5.4e-5/1.4e-6	2.6e-7/2.4e-7	7.7e-11/4.8e-15	5.8e-8/8.7e-12	6.3e-3/2.4e-2	
RMSD (mm) (L/R)	0.620/0.630;	0.910/0.840;	0.710/0.690;	0.661/0.620;	0.650/0.538;	0.542/0.514
	4.7e-1/1.4e-6	1.8e-3/1.1e-6	1.0e-1/4.8e-9	6.2e-3/2.7e-5	8.3e-3/8.0e-3	

 TABLE III. Results of dataset A: Means of the metrics and P values of paired-tests comparing different methods with LLL method for both left and right hippocampus

Bold: value of LLL method is significantly (P < 0.05) better than all other methods.

Dataset A. As shown in Figure 12, all the segmentation methods under study shared a similar pattern, i.e., their performance measured by a leave-one-out validation gradually improved with the increase of the number of atlases used in segmentation. For all the methods, an atlas selection strategy based on NMI similarity metric was used to select atlases [Aljabar et al., 2009; Collins and Pruessner, 2010; Leung et al., 2011; Wolz et al., 2010a; Wu et al., 2007]. Since LWGU and LWINV had similar performance, only LWGU was performed in this experiment.

TABLE IV. Results of dataset B: Means of the metrics and P values of paired-tests comparing the different m	ethods
with LLL method for both side of the hippocampus	

	MV P value (L/R)	STAPLE P value (L/R)	LWGU P value (L/R)	LWINV P value (L/R)	NLP P value (L/R)	LLL
Dice (L/R)	0.872/0.876;	0.830/0.836;	0.887/0.892;	0.891/0.896;	0.898/0.902;	0.907/0.911
	9.1e-11/4.2e-13	5.6e-7/9.0e-8	5.5e-9/5.2e-10	1.1e-7/1.5e-8	8.6e-6/5.1e-5	
Jaccard (L/R)	0.774/0.780;	0.716/0.724;	0.797/0.805;	0.808/0.819;	0.816/0.822;	0.830/0.838
	2.6e-11/3.2e-13	1.3e-7/1.9e-8	2.6e-9/5.0e-10	2.2e-5/2.8e-3	8.5e-6/5.2e-5	
Precision (L/R)	0.898/0.898;	0.737/0.743;	0.912/0.901;	0.915/0.915;	0.911/0.908;	0.918/0.917
	9.5e-4/2.0e-3	1.1e-11/1.1e-12	5.7e-3/4.1e-7	2.3e-1/5.5e-1	1.9e-3/6.5e-4	
Recall (L/R)	0.854/0.861;	0.967/0.969;	0.892/0.889;	0.865/0.873;	0.890/0.899;	0.899/0.908
	2.5e-7/4.4e-9	1.0/1.0	1.9e-5/6.3e-9	8.4e-9/2.1e-10	5.1e-6/2.5e-4	
HD (mm) (L/R)	3.257/3.347;	4.111/3.711;	3.900/3.631;	3.202/3.130;	3.004/3.062;	2.812/0.955
	3.6e-2/1.8e-2	4.9e-5/2.9e-4	1.3e-7/1.6e-3	4.5e-2/1.7e-1	3.7e-1/1.2e-1	
HD95 (mm) (L/R)	1.300/1.389;	1.928/1.763;	1.618/1.573;	1.320/1.360;	1.132/1.069;	1.083/1.041
	3.7e-3/3.7e-5	2.1e-4/3.4e-4	3.4e-8/3.2e-7	1.1e-3/4.2e-4	3.0e-2/2.9e-1	
MD (mm) (L/R)	0.336/0.333;	0.342/0.345;	0.457/0.438;	0.370/0.340;	0.198/0.193;	0.229/0.214
	3.2e-8/2.6e-8	2.7e-13/1.2e-12	3.2e-12/4.7e-12	9.8e-9/8.2e-8	8.6e-5/1.6e-4	
ASSD (mm) (L/R)	0.291/0.287;	0.457/0.435;	0.358/0.343;	0.330/0.320;	0.226/0.217;	0.204/0.198
	6.6e-10/4.0e-11	4.2e-7/1.9e-7	9.6e-14/6.7e-14	1.2e-11/4.7e-12	9.5e-5/9.2e-5	
RMSD (mm) (L/R)	0.593/0.600;	0.821/0.778;	0.688/0.669;	0.570/0.600;	0.509/0.499;	0.478/0.474
	7.3e-9/2.1e-7	3.4e-6/3.5e-6	1.0e-12/6.7e-10	1.1e-5/2.3e-6	1.4e-4/6.1e-4	

Bold: value of LLL method is significantly (P < 0.05) better than all other methods.

	MV P value (L/R)	STAPLE P value (L/R)	LWGU P value (L/R)	LWINV P value (L/R)	NLP P value (L/R)	LLL
Dice (L/R)	0.891/0.888;	0.860/0.853;	0.897/0.891;	0.899/0.900;	0.902/0.895;	0.908/0.907
	1.5e-14/3.0e-4	9.2e-16/2.2e-8	3.7e-10/2.1e-10	4.3e-10/1.3e-2	3.2e-6/1.4e-5	
Jaccard (L/R)	0.804/0.800;	0.757/0.748;	0.814/0.804;	0.817/0.819;	0.821/0.811;	0.833/0.832
	5.9e-15/2.8e-5	1.4e-16/1.9e-9	1.7e-10/2.6e-11	2.0e-10/6.7e-3	2.5e-6/2.9e-6	
Precision (L/R)	0.910/0.910;	0.778/0.770;	0.917/0.924;	0.917/0.923;	0.906/0.907;	0.915/0.926
	1.0e-4/6.3e-5	5.4e-30/5.0e-22	9.3e-1/4.5e-1	9.1e-1/2.1e-1	2.1e-9/4.5e-10	
Recall (L/R)	0.875/0.870;	0.968/0.966;	0.884/0.880;	0.881/0.865;	0.900/0.886;	0.904/0.892
	6.3e-14/2.2e-3	1.0/1.0	2.3e-15/5.9e-2	2.1e-14/3.5e-13	8.7e-3/1.6e-2	
HD (mm) (L/R)	3.035/2.986;	3.551/3.635;	2.950/3.040;	2.960/3.070;	3.000/3.174;	2.850/2.962
	1.4e-4/1.9e-1	3.5e-6/3.1e-6	5.5e-2/9.6e-2	2.2e-3/2.0e-1	1.2e-2/5.9e-3	
HD95 (mm) (L/R)	1.159/1.231;	1.430/1.579;	1.147/1.210;	1.135/1.170;	1.084/1.183;	1.068/1.142
	6.6e-3/5.0e-2	4.2e-7/3.0e-5	8.1e-2/3.6e-2	2.7e-2/3.7e-1	3.8e-1/1.5e-2	
MD (mm) (L/R)	0.287/0.310;	0.302/0.333;	0.240/0.270;	0.254/0.270;	0.213/0.235;	0.228/0.263
	5.1e-9/1.1e-8	4.6e-19/7.7e-12	2.3e-3/6.8e-6	1.7e-5/3.2e-3	1.3e-6/3.8e-7	
ASSD (mm) (L/R)	0.259/0.280;	0.370/0.416;	0.228/0.250;	0.235/0.250;	0.228/0.258;	0.213/0.233
	8.1e-15/7.9e-4	1.2e-17/7.5e-10	6.9e-3/1.2e-3	5.9e-9/4.9e-2	1.3e-5/3.0e-5	
RMSD (mm) (L/R)	0.547/0.574;	0.676/0.735;	0.507/0.526;	0.520/0.536;	0.509/0.546;	0.489/0.513
	6.6e-14/5.9e-4	6.6e-15/6.1e-9	4.6e-2/7.9e-2	3.1e-9/1.9e-2	1.2e-5/1.5e-5	

 TABLE V. Results of dataset C: Means of the metrics and P values of paired-tests comparing the different methods with LLL method for both side of the hippocampus

Bold: value of LLL method is significantly (P < 0.05) better than all other methods.



Figure 9.

Relative volume differences (RVD) (first row) and absolute value of RVD (ARVD) (second row) between the segmentation results of automatic methods and the manual label on the three datasets. Figures at the right side are the zoomed in version of figures at the left side. On each box, the central mark is the median, and edges of the box are the 25th and 75th percentiles. Whiskers extend from each end of the box to the adjacent values in the dataset and the extreme values within I times the interquartile range from the ends of the box. Outliers are data with values beyond the ends of the whiskers. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

	MV;	STAPLE	LWGU;	LWINV;	NLP;	LLL;
	RVD (%);	RVD (%);	RVD (%);	RVD (%);	RVD (%);	RVD (%);
Methods	ARVD (%)	ARVD (%)	ARVD (%)	ARVD (%)	ARVD (%)	ARVD (%)
Dataset A left	5.7 ± 11.5;	$-37.3 \pm 24.7;$	2.2 ± 7.8 ;	$4.8 \pm 9.8;$	$1.9 \pm 8.7;$	1.9 ± 8.6;
correlation	$10.7 \pm 6.8;$	$37.3 \pm 24.7;$	6.7 ± 4.5;	$9.2 \pm 5.5;$	$7.5 \pm 4.6;$	$7.1 \pm 5.1;$
	0.922	0.764	0.956	0.942	0.952	0.952
Dataset A right	$6.0 \pm 12.6;$	$-45.1 \pm 31.1;$	$-2.2 \pm 19.5;$	$4.3 \pm 10.8;$	$2.8 \pm 9.7;$	2.1 ± 9.1 ;
correlation	$11.3 \pm 7.9;$	$45.1 \pm 31.1;$	$11.7 \pm 15.6;$	$9.3 \pm 6.7;$	$8.2 \pm 5.8;$	$7.5 \pm 5.5;$
	0.897	0.716	0.795	0.924	0.931	0.940
Dataset B left	$3.3 \pm 12.2;$	$-33.5 \pm 22.2;$	$4.0 \pm 8.9;$	$4.2 \pm 9.8;$	$0.7 \pm 8.7;$	$0.6 \pm 8.3;$
correlation	$10.5 \pm 6.8;$	$33.5 \pm 22.2;$	$8.1 \pm 5.3;$	$8.9 \pm 5.7;$	$6.7 \pm 5.4;$	$6.3 \pm 5.3;$
	0.918	0.826	0.952	0.945	0.957	0.959
Dataset B Right	$4.1 \pm 13.8;$	$-35.6 \pm 27.9;$	$5.2 \pm 9.9;$	$5.0 \pm 11.1;$	$2.0 \pm 9.4;$	$1.6\pm9.1;$
correlation	$11.5 \pm 8.5;$	$35.6 \pm 27.9;$	$9.4 \pm 5.8;$	$10.0 \pm 6.8;$	$7.7 \pm 5.5;$	$7.4\pm5.3;$
	0.889	0.759	0.937	0.926	0.943	0.944
Dataset C left	$4.0 \pm 8.7;$	$-27.8 \pm 19.9;$	$4.4 \pm 7.5;$	$6.0 \pm 8.7;$	$2.1 \pm 8.4;$	3.3 ± 8.6;
correlation	$7.4 \pm 6.0;$	$27.8 \pm 19.9;$	$7.0 \pm 5.0;$	$8.2 \pm 6.6;$	$6.1 \pm 6.1;$	$6.0\pm6.1;$
	0.923	0.803	0.945	0.914	0.923	0.947
Dataset C right	$3.5 \pm 8.9;$	$-25.9 \pm 15.3;$	$3.3 \pm 8.3;$	$3.6 \pm 8.5;$	$0.3 \pm 8.4;$	$0.8\pm8.1;$
correlation	$7.8 \pm 5.5;$	$25.9 \pm 15.3;$	$7.3 \pm 5.1;$	$7.5 \pm 5.3;$	$6.5 \pm 5.2;$	$6.3\pm5.0;$
	0.932	0.902	0.939	0.937	0.940	0.945

TABLE VI. Relative volume difference (RVD), absolute value of RVD (ARVD), and correlation between manual segmentation and segmentation results obtained by automatic methods

Highest Pearson correlation coefficients are shown in bold.

DISCUSSION AND CONCLUSION

In this study, we propose a local label learning (LLL) strategy for multi-atlas based image segmentation. Instead of explicitly defining a weighting model to fuse the atlas labels, we utilize a machine learning method to build voxel-wise classifiers based on image appearance and texture information. To get a robust classifier that generalizes well for each target voxel, we adopt a local patch strategy to get a training set with abundant appearance and texture information on which an L1-regularized SVM classifier is built in conjunction with a *k*NN training sample selection strategy.

Our method has the following novelties. First, an L1regularized supervised learning method is utilized to learn the relationship between the segmentation label and image appearance/texture for each voxel. The supervised learning method can take rich information as input for learning a mapping from images to the segmentation label, and the adopted L1 SVM can handle the potential redundant information of image features. Besides the feature extraction method adopted in our method, other sophisticated feature extraction techniques can also be adopted in this framework [Dunn and Higgins, 1995]. Second, a local patch strategy is used to get a training set for each voxel to be segmented. Utilizing such a strategy to get the training set not only increases the number of training samples, but also minimizes the partial volume effect due to imaging resolution and the adverse effect of image registration errors. Finally, the kNN strategy based SVM classification

is utilized to get a balanced training dataset. The *k*NN strategy has been demonstrated successful for learning problems with unbalanced training samples [Zhang et al., 2006].

The comparison results have demonstrated that our method could obtain better performance for segmenting images with different spatial resolutions than alternative state-of-the-art methods. It is worth noting that a direct comparison of results across publications is difficult and can be affected by several factors, such as the segmentation protocol, the imaging protocol, and the patient population [Collins and Pruessner, 2010]. Since large labels might lead to larger overlap values in segmentation evaluation [Rohlfing et al., 2004a], multiple segmentation metrics should be used to comprehensively evaluate the segmentation results. One summary of the hippocampal segmentation performance reported in recent literature can be found in Table X of [Wang et al., 2011a]. Most of the recently published results had Dice index less than 0.9 and the method combining multi-atlas-based segmentation with an error correction step achieved the highest Dice index of 0.908 [Wang et al., 2011a]. We got a similar performance with the highest Dice index of 0.910 for images from subjects of mixed diagnostic groups of Alzheimer's disease. Our results have also indicated that the hippocampus segmentation performance is hinged on the spatial resolution of images to be segmented.

One recent study [Heckmann et al., 2011] has provided a publicly available repository of anatomically segmented brain images for the ADNI dataset. The segmentation





Hippocampal segmentation results obtained by different methods. One subject was randomly chosen from each dataset. For each subject, the first row shows the segmentation results produced by different methods, the second row demonstrates their corresponding surface rendering results, and the difference between results of manual and automatic segmentation methods was showed in the third row (red: manual segmentation results, green: automated segmentation results, blue: overlap between manual and automated segmentation results). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

results were also generated by a multi-atlas based segmentation method [Heckemann et al., 2010]. Though hippocampal labels were available for these images, they were obtained with a segmentation protocol different from ours. We did not apply out method to the dataset with hippocampus labels provided by the study [Heckemann et al., 2011] for following reasons. First, our method is not suitable for segmenting hippocampus based on atlases labeled without the hippocampal tail. Since the hippocampal main body and tail have similar intensity information as shown in Figure 13, our method, achieving segmentation based

TABLE VII. Segmentation performance of across different datasets

$\begin{array}{llllllllllllllllllllllllllllllllllll$	Dice index values (L/R)	
	$B \rightarrow A(A \rightarrow A)$ 0.891/0.898(0.887/0.894) $A \rightarrow C(C \rightarrow C)$ 0.879/0.870(0.908/0.907)	$\begin{array}{l} A \to B(B \to B) \\ 0.900/0.908(0.907/0.911) \\ B \to C(C \to C) \\ 0.882/0.875(0.908/0.907) \end{array}$

 $X \rightarrow Y$: images of dataset X were used as atlases to segment images of dataset Y.

Methods	Tu et al. [2008]	Aljabar et al. [2007]	LWGU; <i>P</i> value	LWINV; <i>P</i> value	NLP; P value	LLL
Dice	0.64 ± 0.06	0.75 ± 0.07	0.828 ± 0.031 ; 7.8e-4	0.828 ± 0.031 ; 4.4e-4	0.827 ± 0.030 ; 1.8e-7	0.831 ± 0.030
Jaccard	0.47 ± 0.06	0.60 ± 0.07	0.709 ± 0.043 ; 6.3e-4	0.709 ± 0.043 ; 3.4e-4	0.707 ± 0.042; 1.7e-7	$\textbf{0.714} \pm \textbf{0.042}$
Precision	0.50 ± 0.08	0.77 ± 0.07	0.797 ± 0.053; 2.0e-4	0.797 ± 0.053; 1.6e-4	0.795 ± 0.053; 4.5e-7	$\textbf{0.802} \pm \textbf{0.053}$
Recall	0.92 ± 0.06	0.74 ± 0.10	0.868 ± 0.039; 3.9e-1	0.867 ± 0.039; 2.9e-1	0.866 ± 0.039; 3.2e-2	0.868 ± 0.039
HD (mm)	8.77 ± 4.30	4.45 ± 1.54	5.848 ± 5.590; 1.2e-1	5.844 ± 5.591; 1.4e-1	5.956 ± 5.542; 7.9e-3	$5.800 \pm 5.606;$
HD95 (mm)	4.58 ± 1.68	2.41 ± 0.98	1.851 ± 0.472 ; 9.5e-2	1.851 ± 0.468 ; 6.3e-2	1.840 ± 0.450 ; $3.5e-1$	$1.833 \pm 0.473;$
MD (mm)	1.58 ± 0.50	0.50 ± 0.18	0.351 ± 0.128 ; 7.8e-2	0.351 ± 0.129 ; 5.9e-2	0.342 ± 0.125 ; 9.8e-1	0.347 ± 0.126
ASSD (mm)	1.33 ± 0.33	0.55 ± 0.27	0.339 ± 0.095; 1.1e-3	0.340 ± 0.096 ; 7.2e-4	0.340 ± 0.094 ; 6.7e-5	$0.333 \pm 0.093;$
RMSD (mm)	1.88 ± 0.54	0.95 ± 0.39	0.691 ± 0.165 ; 5.3e-3	0.691 ± 0.164 ; 3.9e-3	0.690 ± 0.160 ; 1.6e-3	0.681 ± 0.161;
RVD (%)	94 ± 37	-4 ± 15	9.7 ± 11.1; 8.1e-4	9.7 ± 11.1; 1.4e-3	9.8 ± 11.1; 3.5e-4	9.0 ± 11.0;
ARVD (%)	/	/	10.6 ± 10.2 ; 8.8e-3	10.6 ± 10.2 ; 1.2e-2	10.7 ± 10.2 ; 6.7e-3	$\textbf{10.1} \pm \textbf{10.0}$

TABLE	VIII. S	egmentation	performance of	on the p	oublicly	available dataset	of ep	ilepsy	patients and	normal	subj	ects
		- A										

Means values with their standard deviations of the metrics are shown. Bold: value of LLL method is significantly (P < 0.05) better than all other methods.

on image intensity information, will generate results including the hippocampal tail. Our method cannot automatically cut the tail. Second, our method might not be able to effectively handle label errors of atlas. As shown in Figure 13, the segmentation results contained errors. Such segmentation label errors also make the performance evaluation complicated.

One major issue of the multi-atlas based image segmentation methods is their high computational cost, mainly due to the image registration. To reduce the computation cost of image registration, one can select a small number of atlases [Aljabar et al., 2009]. As demonstrated in our experiments, a small set of around 20 atlases could lead to a stable segmentation performance. Recent studies have proposed nonlocal patch-based image labeling strategies with linear image registration based atlas alignment [Coupe et al., 2011; Rousseau et al., 2011]. However, such non-local search procedures may increase the computational cost due to the reason that the computational cost of non-local searching in the label fusion step might be higher than the computational cost of non-rigid image registration, as pointed out in [Rousseau et al., 2011]. For speeding up the image registration and subsequently the image segmentation, it might be a good choice to use



Figure 11.

Hippocampal volumes of subjects from three diagnostic groups. On each box, the central mark is the median, and edges of the box are the 25th and 75th percentiles. Whiskers extend from each end of the box to the adjacent values in the dataset and the extreme values within 1 interquartile range from the ends of the box. Outliers are data with values beyond the ends of the whiskers. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

	NC		MCI			AD				
Methods	Volume (mm ³); left; right	Volume (mm ³); left; right	<i>t-</i> Test between MCI and NC (<i>P</i> value)	Effect size (d)	Sample size	Volume (mm ³); left; right	<i>t-</i> Test between AD and NC (<i>P</i> value)	Effect size (d)	Sample size	
Manual segmentation	2,623 ± 393; 2,826 ± 450	2,266 ± 484; 2,482 ± 525	5.3e-3	0.79	25	2,044 ± 586; 2,328 ± 618	9.0e-5	1.07	14	
MV	2,371 ± 315; 2,574 ± 366	2,088 ± 343; 2,317 ± 379	5.0e-3	0.79	25	2,078 ± 474; 2,322 ± 432	8.2e-3	0.70	32	
STAPLE	3,114 ± 343; 3,338 ± 416	2,803 ± 373; 3,087 ± 399	6.2e-3	0.77	26	2,765 ± 531; 3,177 ± 447	1.8e-2	0.62	41	
LWGU	2,417 ± 340; 2,610 ± 384	2,089 ± 386; 2,322 ± 435	3.8e-3	0.82	23	2,055 ± 515; 2,290 ± 492	2.8e-3	0.79	25	
LWINV	2,397 ± 332; 2,593 ± 379	2,084 ± 378; 2,297 ± 427	3.6e-3	0.82	23	2,057 ± 500; 2,235 ± 451	1.3e-3	0.86	21	
NLP	2,506 ± 364; 2,704 ± 403	2,173 ± 407; 2,394 ± 473	4.6e-3	0.80	24	2,103 ± 566; 2,285 ± 503	6.7e-4	0.91	19	
LLL	2,523 ± 369; 2,736 ± 429	2,161 ± 418; 2,399 ± 473	2.8e-3	0.85	22	2,108 ± 552; 2,304 ± 499	5.3e-4	0.93	18	

TABLE IX. Normalized hippocampal volume information of three diagnostic groups estimated by different methods, and two sample t-tests, effect sizes, and sample sizes for different methods to detect a difference in total hippocampal volumes between NC group and MCI group, as well as between NC group and AD group

graphics processing units (GPUs) [Huang et al., 2011; Samant et al., 2008; Sharp et al., 2007] since a GPU-based image registration can achieve a speedup of 25 times for atlas-based brain image segmentation [Han et al., 2009]. In this study, we have implemented our algorithm using Matlab. It took about 7 min to fuse labels for segmenting one side of the hippocampi using single thread on a computation workstation (Intel xeonx5667) with CPUs of 3.07 GHZ.

The proposed method is designed for problems with two labels: foreground or background. However, it is straightforward to use our method in segmentation problems with multiple structures by iteratively segmenting one structure at a time. The method can also be extended for segmentation problems with multiple structures by



Figure 12.

Segmentation performance as a function of the number of atlases used. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

replacing the two-class SVM classifiers with multi-class classifiers, such as random forests [Breiman, 2001]. Furthermore, if fuzzy label models such as those based on distance transforms are used, our method can also be extended by adopting regression techniques.

The proposed method could be further improved using the following strategies. First, shape regularization can be incorporated into the image segmentation framework. In this study, we performed voxel-wise classification. We expect that smoother labels could be obtained by explicitly including shape constrains. Second, our algorithm could be extended to handle multiple brain structures with a multi-class based supervised learning strategy.

ACKNOWLEDGMENTS

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.

NOTE ADDED TO PROOF

Due to a coding error in the authors' program, image features from the second order difference filters (SODs) were not used in the experiments. We have corrected and evaluated the algorithm's segmentation performance on datasets A, B, and C. The results presented in Tables III-V, Table II, and Table VII were updated. The updated segmentation accuracy measures were close to those



Figure 13.

Ten randomly selected images and their segmentation labels of hippocampus obtained from the results provided by the study [Heckemann et al., 2011]. Each row shows one image's two slices and their corresponding segmentation labels. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

presented in the paper. However, the top 10 most frequently selected features were different, and some features from Laplacian filters were replaced with features from SODs since both SOD and Laplacian filters captured edge information and played similar roles in the segmentation.

REFERENCES

- Aljabar P, Heckeman R, Hammers A, Hajnal JV, Rueckert D (2007): Classifier selection strategies for label fusion using large atlas databases. Med Image Comput Comput Assist Interv 4791:523–531.
- Aljabar P, Heckemann RA, Hammers A, Hajnal JV, Rueckert D (2009): Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. Neuroimage 46:726–738.
- Artaechevarria X, Munoz-Barrutia A, Ortiz-de-Solorzano C (2008): Effleient classifier generation and weighted voting for aflasbased segmentation: Two small steps faster and closer to the combination oracle. SPIE Med Imag 2008:6914.
- Artaechevarria X, Munoz-Barrutia A, Ortiz-de-Solorzano C (2009): Combination strategies in multi-atlas image segmentation: Application to brain MR data. IEEE Trans Image Process 28:1266–1277.
- Ashburner J, Friston KJ (2005): Unified segmentation. Neuroimage 26:839–851.
- Asman AJ, Landman BA (2011): Robust statistical label fusion through consensus level, labeler accuracy, and truth estimation (COLLATE). IEEE Trans Image Process 30:1779–1794.
- Asman AJ, Landman BA (2012): Non-local STAPLE: An intensitydriven multi-atlas rater model. Med Image Comput Comput Assist Interv 15:426–434.
- Avants BB, Epstein CL, Grossman M, Gee JC (2008): Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Med Image Anal 12:26–41.
- Avants BB, Yushkevich P, Pluta J, Minkoff D, Korczykowski M, Detre J, Gee JC (2010): The optimal template effect in hippocampus studies of diseased populations. Neuroimage 49:2457–2466.
- Bajcsy R, Lieberson R, Reivich M (1983): A computerized system for the elastic matching of deformed radiographic images to idealized atlas images. J Comput Assist Tomogr 7:618–625.
- Breiman L (2001): Random forests. Mach Learn 45:5-32.
- Collins DL, Holmes CJ, Peters TM, Evans AC (1995): Automatic 3-D model-based neuroanatomical segmentation. Human Brain Mapp 3:190–208.
- Collins DL, Pruessner JC (2010): Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion. Neuroimage 52:1355–1366.
- Collins DL, Zijdenbos AP, Baare WFC, Evans AC (1999): ANI-MAL+INSECT: Improved cortical structure segmentation. Inf Process Med Imaging 1613:210–223.
- Coupe P, Manjon JV, Fonov V, Pruessner J, Robles M, Collins DL (2011): Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. Neuroimage 54:940–954.
- Dunn D, Higgins WE (1995): Optimal Gabor filters for texture segmentation. IEEE Trans Image Process 4:947–964.
- Eng J (2003): Sample size estimation: How many individuals should be studied? Radiology 227:309–313.
- Fan RE, Chang KW, Hsieh CJ, Wang, XR, Lin CJ (2008): LIBLIN-EAR: A library for large linear classification. J Mach Learn Res 9:1871–1874.
- Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S,

Montillo A, Makris N, Rosen B, Dale AM (2002): Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. Neuron 33:341–355.

- Fischl B, van der Kouwe A, Destrieux C, Halgren E, Segonne F, Salat DH, Busa E, Seidman LJ, Goldstein J, Kennedy D, Caviness V, Makris N, Rosen B, Dale AM (2004): Automatically parcellating the human cerebral cortex. Cereb Cortex 14:11–22.
- Gee JC, Reivich M, Bajcsy R (1993): Elastically deforming 3D atlas to match anatomical brain images. J Comput Assist Tomogr 17:225–236.
- Gering DT, Nabavi A, Kikinis R, Hata N, O'Donnell LJ, Grimson WE, Jolesz FA, Black PM, Wells WM III (2001): An integrated visualization system for surgical planning and guidance using image fusion and an open MR. J Magn Reson Imaging 13:967–975.
- Hajnal JV, Xue H, Srinivasan L, Jiang SZ, Rutherford M, Edwards AD, Rueckert D (2007): Automatic segmentation and reconstruction of the cortex from neonatal MRI. Neuroimage 38:461–477.
- Han X, Fischl B (2007): Atlas renormalization for improved brain MR image segmentation across scanner platforms. IEEE Trans Med Imaging 26:479–486.
- Han X, Hibbard LS, Willcut V (2009): GPU-accelerated, gradientfree MI deformable registration for atlas-based MR brain image segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop, Miami, FL, pp 141–148.
- Hao Y, Liu J, Duan Y, Zhang X, Yu C, Jiang T, Fan Y (2012): Local label learning (L3) for multi-atlas based segmentation SPIE Med Imaging 2012:8314.
- Heckemann RA, Hajnal JV, Aljabar P, Rueckert D, Hammers A (2006): Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. Neuroimage 33:115– 126.
- Heckemann RA, Keihaninejad S, Aljabar P, Gray KR, Nielsen C, Rueckert D, Hajnal JV, Hammers A (2011): Automatic morphometry in Alzheimer's disease and mild cognitive impairment. Neuroimage 56:2024–2037.
- Heckemann R, Keihaninejad S, Aljabar P, Rueckert D, Hajnal JV, Hammers A (2010): Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation. Neuroimage 51: 221–227.
- Huang TY, Tang YW, Ju SY (2011): Accelerating image registration of MRI by GPU-based parallel computation. Magn Reson Imaging 29:712–716.
- Iosifescu DV, Shenton ME, Warfield SK, Kikinis R, Dengler J, Jolesz FA, McCarley RW (1997): An automated registration algorithm for measuring MRI subcortical brain structures. Neuroimage 6:13–25.
- Isgum I, Staring M, Rutten A, Prokop M, Viergever MA, van Ginneken B (2009): Multi-atlas-based segmentation with local decision fusion–Application to cardiac and aortic segmentation in CT scans. IEEE Trans Med Imaging 28:1000–1010.
- Jack CR, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, Borowski B, Britson PJ, Whitwell JL, Ward C, Dale AM, Felmlee JP, Gunter JL, Hill DLG, Killiany R, Schuff N, Fox-Bosetti S, Lin C, Studholme C, DeCarli CS, Krueger G, Ward HA, Metzger GJ, Scott KT, Mallozzi R, Blezek D, Levy J, Debbins JP, Fleisher AS, Albert M, Green R, Bartzokis G, Glover G, Mugler J, Weiner MW, Study A (2008): The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J Magn Reson Imaging 27:685–691.

- Jafari-Khouzani K, Elisevich KV, Patel S, Soltanian-Zadeh H (2011): Dataset of magnetic resonance images of nonepileptic subjects and temporal lobe epilepsy patients for validation of hippocampal segmentation techniques. Neuroinformatics 9: 335–346.
- Jia H, Yap PT, Shen D (2012): Iterative multi-atlas-based multiimage segmentation with tree-based registration. Neuroimage 59:422–430.
- Jovicich J, Czanner S, Greve D, Haley E, van der Kouwe A, Gollub R, Kennedy D, Schmitt F, Brown G, MacFall J, Fischl B, Dale A (2006): Reliability in multi-site structural MRI studies: Effects of gradient non-linearity correction on phantom and human data. Neuroimage 30:436–443.
- Khan AR, Cherbuin N, Wen W, Anstey KJ, Sachdev P, Beg MF (2011): Optimal weights for local multi-atlas fusion using supervised learning and dynamic information (SuperDyn): Validation on hippocampus segmentation. Neuroimage 56:126–139.
- Khan AR, Chung MK, Beg MF (2009): Robust atlas-based brain segmentation using multi-structure confidence-weighted registration. Med Image Comput Comput Assist Interv 12:549–557.
- Khan AR, Wang L, Beg MF (2008): FreeSurfer-initiated fullyautomated subcortical brain segmentation in MRI using large deformation diffeomorphic metric mapping. Neuroimage 41: 735–746.
- Kittler J, Hatef M, Duin RPW, Matas J (1998): On combining classifiers. IEEE Trans Pattern Anal Mach Intell 20:226–239.
- Langerak TR, van der Heide UA, Kotte AN, Viergever MA, van Vulpen M, Pluim JP (2010): Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). IEEE Trans Image Process 29: 2000–2008.
- Leung KK, Barnes J, Modat M, Ridgway GR, Bartlett JW, Fox NC, Ourselin S, Initia ADN (2011): Brain MAPS: An automated, accurate and robust brain extraction technique using a template library. Neuroimage 55:1091–1108.
- Leung KK, Barnes J, Ridgway GR, Bartlett JW, Clarkson MJ, Macdonald K, Schuff N, Fox NC, Ourselin S (2010): Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. Neuroimage 51:1345–1359.
- Leventon ME, Grimson WEL, Faugeras O (2000): Statistical shape influence in geodesic active contours. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Volume 1, Hilton Head Island, SC, pp. 316–323.
- Lotjonen JMP, Wolz R, Koikkalainen JR, Thurfjell L, Waldemar G, Soininen H, Rueckert D, Initi ASDN (2010): Fast and robust multi-atlas segmentation of brain magnetic resonance images. Neuroimage 49:2352–2365.
- Magnotta VA, Heckel D, Andreasen NC, Cizadlo T, Corson PW, Ehrhardt JC, Yuh WT (1999): Measurement of brain structures with artificial neural networks: Two- and three-dimensional applications. Radiology 211:781–790.
- Marroquin JL, Santana EA, Botello S (2003) Hidden Markov measure field models for image segmentation. IEEE Trans Pattern Anal Mach Intell 25:1380–1387.
- Mazziotta JC, Valentino D, Grafton S, Bookstein F, Pelizzari C, Chen G, Toga AW (1991): Relating structure to function in vivo with tomographic imaging. Proc Ciba Found Symp 163: 93–112.
- Morra JH, Tu Z, Apostolova LG, Green AE, Avedissian C, Madsen SK, Parikshak N, Toga AW, Jack CR, Schuff N, Weiner MW, Thompson PM, Initi ASDN (2009a): Automated

mapping of hippocampal atrophy in 1-year repeat MRI data from 490 subjects with Alzheimer's disease, mild cognitive impairment, and elderly controls. Neuroimage 45:S3–S15.

- Morra JH, Tu Z, Apostolova LG, Green AE, Toga AW, Thompson PM (2010): Comparison of AdaBoost and support vector machines for detecting Alzheimer's disease through automated hippocampal segmentation. IEEE Trans Med Imaging 29:30–43.
- Morra JH, Tu ZW, Apostolova LG, Green AE, Avedissian C, Madsen SK, Parikshak N, Hua X, Toga AW, Jack CR, Weiner MW, Thompson PM (2009b): Validation of a fully automated 3D hippocampal segmentation method using subjects with Alzheimer's disease, mild cognitive impairment, and elderly controls. Neuroimage 44:1439–1439.
- Ostby Y, Tamnes CK, Fjell AM, Westlye LT, Due-Tonnessen P, Walhovd KB (2009): Heterogeneity in subcortical brain development: A structural magnetic resonance imaging study of brain maturation from 8 to 30 years. J Neurosci 29:11772–11782.
- Pierson R, Corson PW, Sears LL, Alicata D, Magnotta V, Oleary D, Andreasen NC (2002): Manual and semiautomated measurement of cerebellar subregions on MR images. Neuroimage 17:61–76.
- Pohl KM, Fisher J, Grimson WE, Kikinis R, Wells, W.M (2006a): A Bayesian model for joint segmentation and registration. Neuroimage 31:228–239.
- Pohl KM, Fisher J, Shenton M, McCarley RW, Grimson WE, Kikinis R, Wells WM (2006b): Logarithm odds maps for shape representation. Med Image Comput Comput Assist Interv 9: 955–963.
- Powell S, Magnotta VA, Johnson H, Jammalamadaka VK, Pierson R, Andreasen NC (2008): Registration and machine learningbased automated segmentation of subcortical and cerebellar brain structures. Neuroimage 39:238–247.
- Rohlfing T, Brandt R, Menzel R, Maurer CR Jr (2004a): Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. Neuroimage 21:1428–1442.
- Rohlfing T, Maurer CR (2007): Shape-based averaging. IEEE Trans Med Imaging 16:153–161.
- Rohlfing T, Russakoff DB, Maurer CR Jr (2004b): Performancebased classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. IEEE Trans Med Imaging 23:983–994.
- Rousseau F, Habas PA, Studholme C (2011): A supervised patchbased approach for human brain labeling. IEEE Trans Med Imaging 30:1852–1862.
- Sabuncu MR, Yeo BT, Van Leemput K, Fischl B, Golland, P(2010): A generative model for image segmentation based on label fusion. IEEE Trans Med Imaging 29:1714–1729.
- Samant SS, Xia J, Muyan-Ozcelik P, Owens JD (2008): High performance computing for deformable image registration: Towards a new paradigm in adaptive radiotherapy. Med Phys 35:3546–3553.
- Schuff N, Woerner N, Boreta L, Kornfield T, Shaw LM, Trojanowski JQ, Thompson PM, Jack CR, Weiner MW, Initi ASDN (2009): MRI of hippocampal volume loss in early Alzheimers disease in relation to ApoE genotype and biomarkers. Brain 132:1067–1077.
- Sdika M (2010): Combining atlas based segmentation and intensity classification with nearest neighbor transform and accuracy weighted vote. Med Image Anal 14:219–226.
- Sharp GC, Kandasamy N, Singh H, Folkert M (2007): GPU-based streaming architectures for fast cone-beam CT image recon-

struction and demons deformable registration. Phys Med Biol 52:5771–5783.

- Sled JG, Zijdenbos AP, Evans AC (1998): A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging 17:87–97.
- Sowell ER, Trauner DA, Gamst A, Jernigan TL (2002): Development of cortical and subcortical brain structures in childhood and adolescence: A structural MRI study. Dev Med Child Neurol 44:4–16.
- Spinks R, Magnotta VA, Andreasen NC, Albright KC, Ziebell S, Nopoulos P, Cassell M (2002): Manual and automated measurement of the whole thalamus and mediodorsal nucleus using magnetic resonance imaging. Neuroimage 17:631–642.
- Toriwaki J, Yoshida H (2009): Fundamentals of Three-Dimensional Digital Image Processing. London: Springer.
- Tu Z, Narr KL, Dollar P, Dinov I, Thompson PM, Toga AW (2008): Brain anatomical structure segmentation by hybrid discriminative/generative models. IEEE Trans Med Imaging 27: 495–508.
- Tu Z, Toga AW (2007): Towards whole brain segmentation by a hybrid model. Med Image Comput Comput Assist Interv 10:169–177.
- Twining CJ, Cootes T, Marsland S, Petrovic V, Schestowitz R, Taylor CJ (2005): A unified information-theoretic approach to groupwise non-rigid registration and model building. Inf Process Med Imaging 3565:1–14.
- van Rikxoort EM, Isgum I, Arzhaeva Y, Staring M, Klein S, Viergever MA, Pluim JPW, van Ginneken B (2010): Adaptive local multi-atlas segmentation: Application to the heart and the caudate nucleus. Med Image Anal 14:39–49.
- Wang H, Das SR, Suh JW, Altinay M, Pluta J, Craige C, Avants B, Yushkevich PA (2011a): A learning-based wrapper method to correct systematic errors in automatic image segmentation: Consistently improved performance in hippocampus, cortex and brain segmentation. Neuroimage 55:968–985.
- Wang H, Suh JW, Das S, Pluta J, Altinay M, Yushkevich P (2011b): Regression-based label fusion for multi-atlas segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, pp 1113–1120.
- Warfield SK, Zou KH, Wells WM (2004): Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Trans Med Imaging 23:903–921.
- Wolz R, Aljabar P, Hajnal JV, Hammers A, Rueckert D (2010a): LEAP: Learning embeddings for atlas propagation. Neuroimage 49:1316–1325.
- Wolz R, Heckemann RA, Aljabar P, Hajnal JV, Hammers A, Lotjonen J, Rueckert D, Initi ASDN (2010b): Measurement of hippocampal atrophy using 4D graph-cut segmentation: Application to ADNI. Neuroimage 52:109–118.
- Wu MJ, Rosano C, Lopez-Garcia P, Carter CS, Aizenstein HJ (2007): Optimum template selection for atlas-based segmentation. Neuroimage 34:1612–1618.
- Yeo BT, Sabuncu M., Desikan R, Fischl B, Golland P (2008): Effects of registration regularization and atlas sharpness on segmentation accuracy. Med Image Anal 12:603–615.
- Yuan GX, Chang KW, Hsieh CJ, Lin C.J (2010): A comparison of optimization methods and software for large-scale L1-regularized linear classification. J Mach Learn Res 11:3183–3234.
- Zhang H, Berg AC, Maire M, Malik J (2006): SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. Comput Vis Pattern Recognit 2006:2126–2136.