A Learning-Based Wrapper Method to Correct Systematic Errors in Automatic Image Segmentation: Consistently Improved Performance in Hippocampus, Cortex and Brain Segmentation

Hongzhi Wang [a1], Sandhitsu R. Das [a], Jung Wook Suh [a], Murat Altinay[a], John Pluta [a,b], Caryne Craige [a], Brian Avants [a], Paul A. Yushkevich [a] and the Alzheimer's Disease Neuroimaging Initiative[*2]

[a] *Penn Image Computing and Science Laboratory, Departments of Radiology, University of Pennsylvania, Philadelphia, PA, USA*

[b] *Center for Functional Neuroimaging, Departments of Neurology and Radiology, University of Pennsylvania, Philadelphia, PA, USA*

Abstract

We propose a simple but generally applicable approach to improving the accuracy of automatic image segmentation algorithms relative to manual segmentations. The approach is based on the hypothesis that a large fraction of the errors produced by automatic segmentation are systematic, i.e., occur consistently from subject to subject, and serves as a wrapper method around a given host segmentation method. The wrapper method attempts to learn the intensity, spatial and contextual patterns associated with systematic segmentation errors produced by the host method on training data for which manual segmentations are available. The method then attempts to correct such errors in segmentations produced by the host method on new images. One practical use of the proposed wrapper method is to adapt existing segmentation tools, without explicit modification, to imaging data and segmentation protocols that are different from those on which the tools were trained and tuned. An open-source implementation of the proposed wrapper method is provided, and can be applied to a wide range of image segmentation problems.

The wrapper method is evaluated with four host brain MRI segmentation methods: hippocampus segmentation using FreeSurfer (Fischl et al., 2002); hippocampus segmentation using multi-atlas label fusion (Artaechevarria et al., 2009); brain extraction using BET (Smith, 2002); and brain tissue segmentation using FAST (Zhang et al.,

[1]Corresponding author. hongzhiw@mail.med.upenn.edu; Telephone 1-917-349-8786; Fax 1-215-615-3681

[2]Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. ADNI investigators include (complete listing available at http://www.loni.ucla.edu/ADNI/Data/ADNI Authorship List.pdf).

2001). The wrapper method generates 72%, 14%, 29% and 21% fewer erroneously segmented voxels than the respective host segmentation methods. In the hippocampus segmentation experiment with multi-atlas label fusion as the host method, the average Dice overlap between reference segmentations and segmentations produced by the wrapper method is 0.908 for normal controls and 0.893 for patients with mild cognitive impairment. Average Dice overlaps of 0.964, 0.905 and 0.951 are obtained for brain extraction, white matter segmentation and gray matter segmentation, respectively.

Key words: medical image segmentation, error correction, AdaBoost, hippocampal segmentation, brain extraction, brain tissue segmentation.

---

1. Introduction

Accurate automatic segmentation is highly desirable in a large number of neuroimaging applications, given the often prohibitive cost of manual segmentation. Many software tools that address specific segmentation problems are available to today's researcher. However, the end-users of these tools are not always able to achieve the high levels of segmentation accuracy reported by the tool developers. For instance, in (Fischl et al., 2002), the authors of FreeSurfer report average Dice overlap of $\sim 80\%$ between automatic segmentation of the hippocampus and manual segmentations. In (Morra et al., 2009, Pardoe et al., 2009), the users of the same tool achieve only $\sim$70% average overlap with the manual segmentations. There are multiple possible causes for such discrepancies. Firstly, the manual segmentation protocols used by the tool developers and the tool users may be different. The prevailing approach to evaluating segmentation accuracy is to compare automatic segmentation to manual segmentation by one or more experts. However, different experts produce different segmentations, and experts from different centers may use different segmentation protocols or even disagree on the anatomical definitions of the underlying anatomy. This issue has been widely discussed in the literature, and significant advances have been made in deriving a consensus from segmentations by experts with varying degrees of reliability (Warfield et al., 2004). However, even these advances do not address the problem of disagreement in protocols and definitions of anatomy. Thus, the automatic method may be performing just fine on the end-user's data, but the end-user's definition of the ground truth may differ from that of the tool developer. The second possible cause of discrepancy is that modern automatic segmentation methods are largely knowledge-based and incorporate expert knowledge in the form of anatomical shape priors, appearance models, and other parameters. This knowledge is often constructed based on some specific dataset that may be consistently different from the end-user's imaging data.

The aim of this paper is to increase the accuracy of existing automatic segmentation methods when applied to end-users' data and evaluated against end-users' manual segmentations. One way to achieve this would be for the end-user to retrain and retune the automatic segmentation method on his or her own data, using his or her own segmentation protocol. This approach is not universally available, and may require scientific and technical expertise far beyond the level needed to apply the segmentation method. We advocate a simpler alternative approach that works with out-of-the-box automatic segmentation software.

When evaluated with respect to manual segmentation, the errors produced by a segmentation algorithm can be categorized into two classes: random errors and consistent errors (Warfield et al., 2008, Aljabar et al., 2009). The random errors are caused by random effects, e.g. image noise and random anatomical variation. They can be reduced by label fusion techniques that combine information from multiple segmentation attempts performed independently, e.g. (Rohlfing et al., 2005, Warfield et al., 2004, Heckemann et al., 2006, Aljabar et al., 2009, Sabuncu et al., 2010). In this paper, we focus on addressing the other type of errors, consistent errors. Consistent errors are errors that follow a systematic pattern and are caused by mistranslating manual segmentation protocols into the criteria followed by the automatic segmentation method. For example, in hippocampal segmentation, an automatic segmentation method may mistakenly classify pockets of cerebrospinal fluid (CSF) inside the hippocampus as parts of the hippocampus, perhaps due to regularization priors and partial volume effects. A human rater may be more likely to classify these pockets as not belonging to the hippocampus. Such a difference between automatic and manual segmentations is systematic because it occurs consistently under a given set of conditions (i.e. low intensity values in a portion of the hippocampus where CSF pockets tend to form). In this paper, we hypothesize that it is feasible for machine learning techniques to learn the conditions under which consistent errors in automatic segmentation occur and to subsequently detect and correct these consistent errors for other images. For example in the hippocampal CSF example given above, a classifier could be built to recognize spatial locations and intensity patterns under which mislabeling of CSF as hippocampal tissue is likely to occur; applying such a classifier to the results of automatic segmentation on a new image could properly relabel some voxels as CSF.

In what follows, we use four different neuroimaging segmentation experiments to demonstrate that the consistent errors between automatic segmentation results and manual segmentations tend to be associated with a consistent pattern of intensity and contextual features, which can be modeled and learned by a machine learning algorithm. The main contribution of our paper is a *wrapper algorithm* that (1) learns these patterns using example manual segmentations and imaging data provided by the end-user; and (2) applies a correction to the automatic segmentation results produced by the out-of-the-box method on the end-user's data. The wrapper method is capable of correcting even very large consistent errors. For example, in an experiment that analyzes FreeSurfer hippocampus segmentation, average Dice overlap between automatic and our reference segmentations is improved from 66% to 84% using only ten reference segmentations for training.

The wrapper method can also be incorporated into the segmentation tool itself, in some cases requiring no additional training data. This scenario is illustrated below in the context of multi-atlas hippocampus segmentation, where the wrapper method provides a small but significant boost to segmentation accuracy, generating highly competitive results for manual/automatic segmentation agreement reported in the literature. Our wrapper method can be easily implemented, and a reference implementation is provided as open-source software.[3]

[3]Source code and documentation at http://www.nitrc.org/projects/segadapter

This paper is organized as follows. The Methods and Materials section summarizes the imaging datasets used to evaluate the proposed approach; details the proposed learning-based wrapper method for segmentation error correction, as well as two variants used to demonstrate the effectiveness of the specific components of the method; and describes the four host automatic segmentation methods to which the wrapper method is applied. The Results section evaluates the performance of wrapper method relative to manual segmentation, compares the variants of the wrapper method, and evaluates the sensitivity of the wrapper method to parameters such as training set size. The strengths, weaknesses and potential use cases of the method are discussed in the Discussion section.

2. Materials and Methods

2.1. Subjects and Imaging

Our study was conducted on three different segmentation problems: segmentation of the hippocampus, brain extraction and brain tissue segmentation. Hippocampus segmentation experiments use the data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) available at http://www.loni.ucla.edu/ADNI. Brain extraction and brain tissue segmentation experiments use the data from the Internet Brain Segmentation Repository (IBSR) provided by the Center for Morphometric Analysis at Massachusetts General Hospital and available at http://www.cma.mgh.harvard.edu/ibsr.

2.1.1. ADNI Data

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.

ADNI MRI data includes structural 1.5T MRI from all 800 subjects and 3T structural MRI from 200 subjects. Our experiments were conducted using a subset of the 3T images for which reference segmentations were already available from an earlier study of hippocampal atrophy rate estimation in MCI patients (Yushkevich et al., 2010). This subset consists of baseline 3T MRI scans of 82 MCI patients and 57 controls.

The reference segmentations for evaluation were generated for the ADNI subjects as follows. First, we applied a landmark-guided atlas-based automatic segmentation method (Pluta et al., 2009) to obtain an initial hippocampal segmentation for each image. This method requires six manually-placed landmarks as input. Each initial segmentation was then edited in three dimensions by M.A. in ITK-SNAP software (Yushkevich et al., 2006a) following a previously validated protocol (Hasboun et al., 1996). The reliability of this segmentation protocol is summarized in the Appendix.

2.1.2. IBSR Data

The dataset contains 18 T1-weighted MR brain images and their manual segmentations. The images provided by IBSR have been normalized into the Talairach space (rotation only) and preprocessed by intensity inhomogeneity correction routines. The images have slice thickness of 1.5 mm with in-plane resolution varying between $1.0\,\mathrm{mm} \times 1.0\,\mathrm{mm}$ and $0.84\,\mathrm{mm} \times 0.84\,\mathrm{mm}$. The manual segmentations contain labels for gray matter, white matter and the ventricles. Notably, cerebrospinal fluid (CSF) outside of the ventricles is assigned the gray matter label in the IBSR segmentations.

2.2. Learning-Based Wrapper Methods

To improve the segmentation results produced by a given host automatic segmentation method, we propose two learning-based wrapper algorithms: a wrapper algorithm with *explicit error correction (EC)* and a wrapper algorithm with *implicit error correction*. Explicit EC explicitly searches for voxels mislabeled by the host method and assigns a new label to them. By contrast, implicit EC assigns new labels to voxels without first determining if they are mislabeled or not. These two methods are equivalent for binary segmentation problems, but for multi-label segmentation, the explicit EC method is more computationally efficient, as explained later in this section.

The data used to train our wrapper method consist of a set of images, a set of corresponding manual segmentations of the structure of interest, and a set of corresponding automatic segmentation results produced by the host method. For evaluation, the wrapper method is applied to test data, consisting of a set of images and corresponding segmentations by the host method. To quantify the performance of our error correction algorithms, corrected and uncorrected automatic segmentations were assessed by comparison with manual segmentations. Since manual segmentations are available for all the subjects in our datasets, for each experiment we randomly partition each dataset into training and test subsets. As a means of cross-validation, experiments are repeated for multiple random partitions.

Next, we describe the error correction wrapper algorithms in detail.

2.2.1. Explicit Error Correction

Fig. 1 summarizes the explicit EC wrapper method. Using this method, a target image is segmented as follows. First, the host segmentation method is used to obtain the initial segmentation of the target image. Second, a classifier attempts to identify voxels that have been mislabeled by the host method. We refer to this as the *error detection* classifier. Next, a second classifier is used to reassign labels to the voxels tagged as mislabeled by the error detection step. This classifier is referred to as *error correction*. To construct the error detection and error correction classifiers, we use training data for which manual segmentations are available and initial segmentations are obtained by running the host method. We now describe the training and application of the error detection and error correction classifiers in detail.

Error Detection as a Binary Classification Problem. Given segmentation results produced by a host segmentation method, our first step is to identify which voxels are mislabeled with respect to the manual segmentations. We formulate this problem as a classification problem, which is addressed via machine learning as follows.

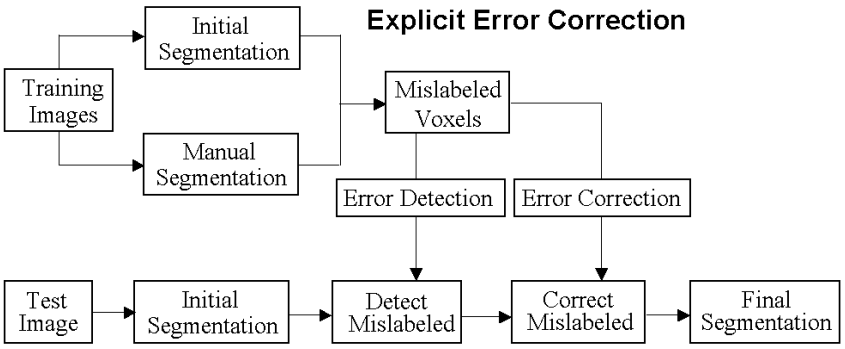

Figure 1: Flow chart of the explicit EC method. The error detection step finds the voxels that are likely to be mislabeled by the host method. The error correction step re-assigns a new label to them. Both error detection and error correction are automatically learned using training segmentations produced by the host method.

We assume that the segmentation problem involves assigning one or more foreground labels to the structures of interest and a background label to the rest of the image. For each foreground label L, we train one classifier to separate correctly labeled voxels from the mislabeled voxels. All voxels across all training images that are assigned the label L by the host segmentation method serve as examples for training the classifier for label L (see Table 2). The features used for training these classifiers are derived from the neighborhood of each voxel, and are discussed later in this section.

The approach is slightly different for the background label, since in many segmentation problems the background region is much larger than the foreground. As before, a classifier is trained to identify voxels incorrectly assigned the background label by the host method. However, when the host method works reasonably well, most voxels incorrectly labeled as background should be in the close proximity of the voxels labeled as foreground. Hence, instead of using all voxels labeled as background by the host method as training examples, we only use the subset of these voxels that lie in a region of interest (ROI) obtained by dilating the set of voxels assigned the foreground label by the host segmentation. In the rest of the paper, we refer to this region of interest as the *working ROI*. The restriction of training to the working ROI excludes the vast majority of irrelevant background voxels from consideration and simplifies the learning problem considerably. In our experiments, we choose the dilation radius such that, in the training set, the working ROIs cover the vast majority of the voxels assigned the foreground label by the manual segmentation (see the Results section for some examples). In certain segmentation problems, the foreground region is provided as input, and there is no need to train the background error detection classifier. For example, in our experiments with brain tissue segmentation, manual brain extraction masks from IBSR define the foreground region, and segmentation does not involve the background label.

Note that for multi-label segmentation problems, the error detection classifiers for different labels perform different classification tasks, i.e., detecting voxels where the given label L has been assigned erroneously by the host method. Learning these tasks

separately decomposes the complex multi-class classification problem into several simpler binary classification problems. However, for segmentation problems with only two labels, the error detection classifiers for the foreground and background labels perform equivalent tasks. Hence, to increase the robustness against overfitting for binary segmentation problems, we train a single error detection classifier for both foreground and background using all voxels within the working ROI.

Error detection classifiers are constructed using the AdaBoost algorithm (Freund and Schapire, 1995), which has shown excellent ability to learn complex patterns in the context of medical image segmentation, as exemplified by the work of Tu et al. (2007) and Morra et al. (2008). AdaBoost builds strong classifiers by combining complementary weak classifiers. Intuitively, AdaBoost iteratively updates the weights associated with each training sample based on the selected weak classifiers, such that the samples that are incorrectly classified receive higher weights. By doing so, weak classifiers selected later in the course of the training complement the previously selected weak classifiers, in the sense that they only focus on classifying samples that have been previously classified incorrectly. Combining these complementary weak classifiers produces a strong classifier that performs better than any single weak classifier.

Following common practice (Viola and Jones, 2001, Tu et al., 2007), we build weak classifiers based on features that are extracted from local image appearance. Let $A^{\Delta x,\Delta y,\Delta z}(i) = I(x_i + \Delta x, y_i + \Delta y, z_i + \Delta z) - \overline{I}$ be the appearance feature at the relative location $(\Delta x, \Delta y, \Delta z)$ for the voxel i with coordinates (x_i, y_i, z_i) . I is the image intensity. To compensate for different intensity ranges across different images, the intensity features are normalized by the mean intensity $\bar{I}$ of the working ROI. Note that our set of appearance features for each voxel includes the complete local image patch, rather than higher-order quantities derived from partial derivatives of image intensity. By including all intensity information in the neighborhood of a voxel, we rely on AdaBoost to find a combination of these intensities that is most useful in the context of classification. More robust features with scale and rotation invariance could be included as well. However, since the brain images used in our experiments have similar scales and orientations, simple appearance features are sufficient to demonstrate the usefulness of our method.

Image appearance features are capable of capturing certain image intensity patterns associated with consistent segmentation errors (e.g., if the host method makes consistent errors in CSF regions, which have relatively darker intensity). However, segmentation errors can also correlate with patterns not captured by the local intensity patch. For example, the host method may consistently overextend its segmentation past a certain anatomical boundary for which there is little intensity contrast. To allow the classifier to learn such patterns of mis-segmentation, we include as features the segmentation labels produced by the host method in the neighborhood of each voxel. We represent these features by $L^{\Delta x,\Delta y,\Delta z}(i) = s(x_i + \Delta x, y_i + \Delta y, z_i + \Delta z)$, where s is the segmentation produced by the host method. We refer to these features as *contextual features* because they supplement intensity features with important contextual information. For instance, they allow the classifier to treat the same intensity patch differently depending on whether it occurs on the boundary of the host segmentation or on the interior. As our experiments below demonstrate, the contextual features are crucially important for the performance of the wrapper method.

To include spatial information, we add the coordinate feature $S_x(i) = x_i - \overline{x}$, $S_y(i) = y_i - \overline{y}$ and $S_z(i) = z_i - \overline{z}$, where $\overline{x}$, $\overline{y}$, $\overline{z}$ are the coordinates of the center of mass of the working ROI. To enhance the spatial correlation, we include the joint feature obtained by multiplying each spatial feature with each appearance and contextual feature. For example, the joint features of appearance and location include $A^{\Delta x,\Delta y,\Delta z}(i)S_x(i)$, $A^{\Delta x,\Delta y,\Delta z}(i)S_y(i)$, and $A^{\Delta x,\Delta y,\Delta z}(i)S_z(i)$. In our experiment, all features are sampled in a $5 \times 5 \times 5$ neighborhood of a given voxel (i.e., $\Delta x, \Delta y, \Delta x \in [-2, 2]$), which yields a total of 1003 features.

Given the response of a feature at each training voxel, e.g. $A^{(0,0,0)}(i)$, we follow (Viola and Jones, 2001) and construct a weak classifier via a linear transform, i.e. $h(A^{(0,0,0)}(i)) = \text{sign}(aA^{(0,0,0)}(i) - b)$, where $a \in \{-1, 1\}$ and b is a threshold. Both parameters are optimized through a linear search such that the weighted misclassification rate is minimized. After the weak classifiers are built, we apply AdaBoost to select and combine them into a single strong classifier. In our experiments, we train every AdaBoost classifier for 500 iterations.

Applying error detection classifiers to test images involves computing the initial segmentation of the test image using the host method, deriving a working ROI for the test image by applying dilation to the initial segmentation, and, for each label L, applying the AdaBoost classifier corresponding to L at each voxel assigned the label L in the initial segmentation. This results in a subset of voxels in the working ROI being marked as mislabeled. These voxels are used as the input for the error-correction classifier.

Error Correction Classifiers. We seek to assign a new, hopefully correct label to the voxels marked mislabeled by the error detection classifiers. For segmentations with only two labels, this step is unnecessary, since correction simply involves flipping the label of the voxels marked as mislabeled by the error correction.

For segmentation problems with more than two labels, we formulate error correction as a multi-class classification problem. Using all voxels incorrectly segmented by the host segmentation method as training exemplars, we train a separate classifier for each label L. Each classifier is trained to separate voxels assigned label L by the manual segmentation from voxels assigned all other labels by the manual segmentation (see Table 2). Again, we use AdaBoost learning with the same set of features described above. As in the case of error detection, we use the mean intensity and the center of mass of the working ROI to normalize the appearance and the spatial features used for error correction.

To assign a new label to a voxel marked mislabeled by error detection, we apply each error correction classifier to that voxel and assign the label whose corresponding classifier gives the strongest response.

2.2.2. Implicit Error Correction

In explicit EC, we explicitly perform error detection and error correction as separate steps. This strategy is efficient because only the potentially mislabeled voxels need to be relabeled for error correction. We also examine an alternative learning-based approach, which we call implicit error correction. In this approach, we skip the error detection step and directly perform error correction upon the initial segmentation. This

method is equivalent to explicit EC where *every voxel* in the working ROI is marked as mislabeled.

Since this learning algorithm aims at directly transferring the segmentation produced by a host method to the corresponding manual segmentation, it implicitly corrects the errors produced by the host method. Fig. 2 summarizes our implicit EC method.

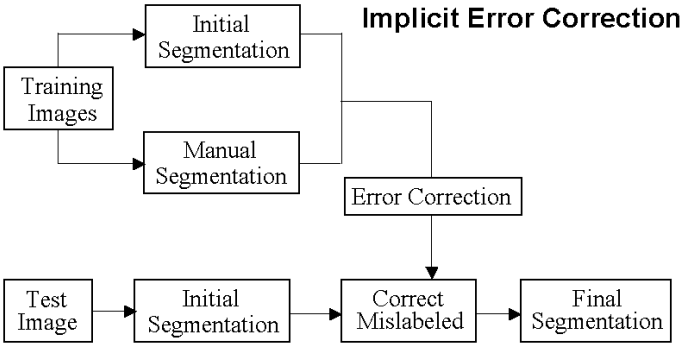

Figure 2: Flow chart of the implicit error correction method.

To train implicit EC classifiers, we use a working ROI obtained by dilating the set of voxels labeled as foreground by the host method. Again, dilation is necessary only when the background label needs to be corrected. Using all voxels within the working ROI, we train one classifier for each label to recognize voxels actually assigned to this label by manual segmentation (see Table 2). It is easy to see that explicit EC is equivalent to implicit EC on segmentation problems with only two labels.

Since implicit EC reevaluates every voxel in the working ROI, it is not affected by the errors produced by the error detection classifiers in explicit EC. This becomes an advantage when error detection classifiers are unreliable. On the other hand, implicit EC has higher computational complexity than explicit EC for both training and testing because implicit EC trains multiple classifiers using all voxels in the working ROI as training examples, while explicit EC trains only a single error detection classifier using the whole working ROI, while using only the voxels marked mislabeled to train the per-label error correction classifiers. The computational complexity of the two methods is compared in Table 1.

2.2.3. Direct Learning

To demonstrate the usefulness of including contextual features (i.e, features derived from the segmentation produced by the host method) in EC classifier learning, we compare our error correction wrapper methods with a variant of implicit EC, where the contextual features are not included as a feature. We call this variant the direct learning (DL) approach. Using the same training data used for implicit EC, we apply AdaBoost to train one DL classifier for each label to recognize voxels manually assigned to this label. The only difference from implicit EC is that we only use appearance and spatial features for learning DL classifiers, while the features $L^{\Delta x,\Delta y,\Delta z}$ are not included.

In principle, in the absence of a segmentation result by a host method, DL should use all voxels in the whole image for training. However, to highlight the contribution of contextual features for learning, we train DL classifiers over the same working ROI as in explicit/implicit EC. Hence, in our experiments, DL partially benefits from the results produced by the host segmentation methods.

The general features and the computational cost of the explicit EC, implicit EC and DL methods are compared in Table 1. The differences in the way the classifiers used by these three methods are specified in Table 2.

	explicit EC	implicit EC	DL
Explicit search for mislabeled voxels	yes	no	no
Contextual features used	yes	yes	no
Computational cost	$(1 + rn_L)N\,(N_A + N_L)$	$n_L N\,(N_A + N_L)$	$n_L N N_A$

Table 1: Summary of explicit EC, implicit EC and DL. n_L is the number of labels. N is the size of the working ROI, in voxels. r is the fraction of voxels mislabeled by the host method. N_A and N_L are the number of appearance features and label features, respectively. Explicit EC usually has a smaller computational cost than implicit EC and DL for multi-label segmentation problems. Implicit EC and explicit EC are equivalent for two-label problems.

2.3. Host Segmentation Methods

The variants of the wrapper method are evaluated in three common MRI segmentation problems using four host methods. For the problem of hippocampus segmentation in ADNI MRI, the host methods are FreeSurfer (Fischl et al., 2002) and a multi-atlas label fusion segmentation approach (Artaechevarria et al., 2009, Sabuncu et al., 2010). For brain extraction in IBSR data, the host method is the Brain Extraction Tool (BET) (Smith, 2002). For the problem of three-tissue segmentation, the FSL FAST algorithm (Zhang et al., 2001) serves as the host method.

	Explicit EC		Implicit EC and DL
	Error Detection	Error Correction	
Number of classifiers	n_L	n_L	n_L
Training exemplars of class 1 for classifier L	$\{i : s(i) = L, m(i) = L\}$	$\{i : s(i) \neq m(i), m(i) = L\}$	$\{i : m(i) = L\}$
Training exemplars for class 0 for classifier L	$\{i : s(i) = L, m(i) \neq L\}$	$\{i : s(i) \neq m(i), m(i) \neq L\}$	$\{i : m(i) \neq L\}$
Voxels in test image to which classifier L is applied	$\{i : s(i) = L\}$	Voxels assigned class 0 by error detection	All voxels in working ROI

Table 2: A comparison of how the explicit and implicit error correction methods train and apply classifiers. Above, n_L denotes the number of labels in a segmentation problem; i indexes voxels in an image; $s(i)$ is the initial segmentation produced by the host method at voxel i; and $m(i)$ is the manual segmentation of voxel i.

Hippocampus segmentation and brain extraction are binary segmentation problems (hippocampus segmentation involves segmenting left and right hippocampi, but these segmentations are performed as independent binary problems). Thus, for these problems the explicit and implicit variants of the EC method are equivalent. The three-tissue segmentation problem involves multiple labels and allows these variants to be compared.

2.3.1. FreeSurfer

FreeSurfer is a software pipeline for the study of cortical and subcortical anatomy. It contains preprocessing components that extract the brain and compensate for intensity inhomogeneity; segmentation tools; and other utilities for cortical and subcortical morphometry. Subcortical segmentation is achieved by aligning the target image with an atlas constructed from a set of manually labeled training images. Although FreeSurfer is not a specialized tool for hippocampus segmentation, due to its popularity and its good general segmentation performance, hippocampal segmentation results by FreeSurfer have been used as a benchmark for evaluating the performance of automatic hippocampal segmentation methods in the recent literature (Morra et al., 2009, Akhondi-Asl et al., 2010, Sanchez-Benavides et al., 2010). As in these papers, we apply FreeSurfer to imaging data with different acquisition parameters from those on which FreeSurfer was trained, and evaluate it against reference segmentations generated using a different hippocampus segmentation protocol. The intention of the FreeSurfer experiment is to demonstrate that the wrapper method can help reconcile these differences in imaging and segmentation protocols, making FreeSurfer perform very well on our data.

In this test, FreeSurfer was applied with the default parameters to segment the left and right hippocampus in each image in the ADNI dataset. 10 cross-validation experiments were performed, with 70 subjects selected at random to form the training set and the remaining 69 subjects forming the test set. Additional experiments with training sets of size 1 to 5, 10 and 20 were also performed to assess the relationship between the size of the training set and the improvement achieved by the wrapper method.

2.3.2. Multi-Atlas Label Fusion

Multi-atlas based segmentation labels a target image by computing one-to-one correspondences with a set of labeled atlases, i.e., images with similar appearance in which the segmentation of the structure of interest is given (Rohlfing et al., 2005). Correspondences are computed using deformable image registration, and segmentation labels are mapped from the coordinate spaces of the different atlases into the coordinate space of the target image. These warped segmentations are combined into a single consensus segmentation using a label fusion strategy. Various fusion strategies have been proposed, majority voting being the simplest. Recent work has demonstrated the effectiveness of strategies where the contribution of each atlas to the consensus segmentation is weighted by the local intensity similarity of the atlas to the target image (Artaechevarria et al., 2009). Because of its simplicity and good performance, multi-atlas segmentation has become a popular approach for medical image segmentation.

Compared with the FreeSurfer experiment, the experiment using multi-atlas label fusion as the host method is designed to demonstrate the contribution of the wrapper

method in the absence of systematic differences due to imaging and segmentation protocols. Thus, the test images and the atlas images in this experiment all come from the ADNI dataset. The multi-atlas experiment is also an example of a scenario where using the wrapper method to improve segmentation performance does not require additional training data beyond that already used by the host algorithm, because the training of the EC classifiers is performed among the atlases in a leave-one-out fashion.

As before, 10 cross-validation experiments were performed. In each experiment, 20 subjects were randomly chosen as atlases and 20 more were chosen as test images. Each atlas was registered to each test image, as well as to each other atlas. Global registration was performed using the FSL FLIRT tool (Smith et al., 2004) with six degrees of freedom and using the default parameters (normalized mutual information similarity metric; search range from -5 to 5 in x, y and z). Deformable registration was performed using the ANTS Symmetric Normalization (SyN) algorithm (Avants et al., 2008), with the cross-correlation similarity metric (with radius 2) and a Gaussian regularizer with $\sigma = 3$. After registration, reference segmentations from each of the atlases were warped into the target image space.

To compute the consensus segmentation of each target image, we use the label fusion strategy determined to be most effective in the recent studies by Artaechevarria et al. (2009) and Sabuncu et al. (2010). Let T_F be a test image and $A^1 = (A^1_F, A^1_S), ..., A^n = (A^n_F, A^n_S)$ be n atlases registered to T_F, with A^i_F denoting the warped atlas image, and A^i_S denoting the corresponding warped reference segmentation. The locally weighted label fusion strategy produces the final segmentation $\hat{T}_S(x)$ as follows:

$$\hat{T}_S(x) = \text{argmax}_{L \in \{1...n_L\}} \sum_{i=1}^{n} w^i(x) \delta(A^i_S(x), L) \tag{1}$$

where L indexes through the labels, n_L is the number of labels (in our case, 2), and δ is the Kronecker delta function. The spatially varying weight $w^i(x)$ measures the confidence that atlas i produces the correct label for the test image at x, which is estimated from the appearance similarity between the test image and the registered atlas image in the neighborhood of x. We apply the summed square distance (SSD) and a Gaussian model (Sabuncu et al., 2010) to estimate the weights as follows:

$$w^i(x) = \frac{exp(-\sum_{j \in \mathcal{N}(x)} \left[T_F(j) - A^i_F(j)\right]^2 / \sigma)}{\sum_{k=1}^{n} w^k(x)} \tag{2}$$

where $\mathcal{N}$ denotes a neighborhood centered at x. We use a $(2r+1) \times (2r+1) \times (2r+1)$ cube-shaped neighborhood specified by the radius r, which is a parameter of the method. To account for absolute intensity differences between the atlases and the target image, instead of using the raw image intensities to estimate the similarity-based weights, we normalize the intensity vector obtained from each local image intensity patch, such that the normalized vector has zero mean and unit variance. To reduce the effects of noise, we spatially smooth the weights for each atlas by a mean filter of the same size as the neighborhood $\mathcal{N}$. After smoothing, the weights are renormalized such that for any x, $\sum_{i=1}^{n} w^i(x) = 1$.

For each cross-validation experiment, this approach generates a consensus segmentation of each test image, as well as a consensus segmentation of each atlas image

by all the remaining atlases. The label fusion approach has two free parameters, the neighborhood radius r and the standard deviation σ in the Gaussian model (2). For each cross-validation experiment, we determine the optimal values of these parameters using the atlas subset in a leave-one-out strategy. That is, we measure the average overlap between the consensus segmentation of each atlas via the remaining atlases and the reference segmentation of that atlas, and find the value of r or σ that maximize this average overlap. Each parameter is optimized separately by evaluating a range of values ($r \in \{1, 2, 3\}$; $\sigma \in \{0.05, 0.1, 0.15, \ldots, 1\}$). Importantly, the reference segmentations of the test images in each cross-validation experiment are not used for finding the optimal parameters r and σ for that experiment, eliminating the possibility of overfitting.

The input to the EC training consists of the atlas images, their consensus segmentations by the remaining atlases (playing the role of the host method segmentation result), and their reference segmentations. To boost the size of the training set, the flipped mirror images of the right hippocampi are combined with the left hippocampi to train the EC classifiers. EC is then applied to the test images and their consensus segmentations. For right hippocampus segmentation, the test images are also mirror flipped before applying the EC classifiers.

2.3.3. Brain Extraction Tool (BET)

BET (Smith, 2002) uses a deformable model to separate the brain from other tissues in MR images. In our experiments, BET was applied with the default parameters to segment each of the 18 brain images from IBSR. The EC method was used to improve the accuracy of brain extraction relative to the brain masks in IBSR. 10 cross-validation experiments were performed. For each cross-validation evaluation, 9 subjects were randomly selected for training the EC method, and the remaining 9 for testing. The brain volumes have millions of voxels, posing a challenge for the AdaBoost learning, which requires loading all data in memory for efficient learning. For efficiency, we randomly selected 1% voxels uniformly from the working ROIs for training.

2.3.4. FMRIB's Automated Segmentation Tool (FAST)

The FAST algorithm (Zhang et al., 2001) is used to segment brain MRI into gray matter, white matter, and CSF. It takes an expectation-maximization strategy to iteratively search for the optimal segmentation and the optimal inhomogeneity bias field correction solution. The solutions are spatially regularized by a Markov Random Field prior to reduce the effects of image noise. In our experiment, the FAST algorithm was applied with the default parameter settings for all 18 subjects. The region of interest for the three tissue segmentation was restricted to the brain by providing the manually computed brain masks in the IBSR dataset as input to FAST. The explicit and implicit versions of the EC method were evaluated in 10 cross-validation experiments with the same partitioning of subjects into training and test sets as in the BET experiment.

3. Results

3.1. FreeSurfer

FreeSurfer hippocampus segmentations tended to be substantially larger than the corresponding reference segmentations, and it was sufficient to use a single-voxel di-

lation to obtain the working ROI for the learning algorithms. On average, this ROI covered 99.7% of the foreground voxels in the reference hippocampus segmentations in the training data. Using this working ROI definition, our implementation of the EC method[4] completed AdaBoost training in 6 hours on a 3 GHZ CPU for each cross-validation experiment. Applying the trained EC classifiers to correct the segmentation for a test image only took a few seconds of CPU time.

	LEFT			RIGHT		
Exp.	initial(Dice)	DL(Dice)	EC(Dice)	initial(Dice)	DL(Dice)	EC(Dice)
1	0.665±0.045	0.838±0.037	**0.864**±0.033	0.658±0.041	0.845±0.030	**0.866**±0.025
2	0.666±0.045	0.839±0.035	**0.863**±0.028	0.651±0.045	0.843±0.031	**0.865**±0.025
3	0.662±0.042	0.837±0.036	**0.861**±0.034	0.655±0.040	0.846±0.028	**0.866**±0.030
4	0.663±0.046	0.833±0.037	**0.861**±0.034	0.661±0.037	0.842±0.026	**0.870**±0.021
5	0.666±0.045	0.839±0.033	**0.865**±0.034	0.655±0.043	0.843±0.031	**0.866**±0.031
6	0.668±0.045	0.838±0.031	**0.864**±0.030	0.655±0.041	0.843±0.026	**0.867**±0.024
7	0.664±0.047	0.843±0.030	**0.865**±0.030	0.648±0.043	0.842±0.031	**0.866**±0.030
8	0.668±0.044	0.838±0.036	**0.863**±0.033	0.659±0.040	0.842±0.031	**0.865**±0.031
9	0.665±0.045	0.842±0.034	**0.867**±0.032	0.652±0.039	0.845±0.032	**0.867**±0.032
10	0.665±0.045	0.840±0.035	**0.865**±0.033	0.656±0.042	0.842±0.031	**0.863**±0.033

Table 3: Results of automatic hippocampus segmentation using FreeSurfer and the wrapper method. Each row gives the average Dice overlap between automatic and reference segmentations for one cross-validation experiment. The bold font highlights the best results.

The average size of the working ROI was 5978 voxels, and the average number of voxels in the reference segmentations was 1598. On average, FreeSurfer produced 1489 mislabeled voxels for each hippocampus. The EC method produced 72.0% fewer errors (418 mislabeled voxels) than FreeSurfer. By contrast, the DL method produced slightly worse results with 472 mislabeled voxels. Table 3 shows the results in terms of Dice overlap with reference segmentations for each of the 10 cross-validation experiments. On average, the EC method increased Dice overlap from 0.660 to 0.865.

Visualization of spatially consistent segmentation errors. Fig. 3 shows examples of the differences between FreeSurfer hippocampus segmentations and reference segmentations (middle column), and the differences after applying the EC wrapper method (right column). These differences appear to follow a consistent spatial pattern, with the FreeSurfer segmentation extending farther in the superior direction than the reference segmentation. Furthermore, FreeSurfer segmentations include white matter structures such as the alveus, as well as some CSF voxels, while the reference segmentations exclude them. Thus, the differences between reference segmentations and FreeSurfer segmentations are associated with specific spatial locations and specific intensity patterns; both of these can be learned easily by a machine learning algorithm, which

[4]Recall that for binary segmentation problems the explicit and implicit EC methods are equivalent.

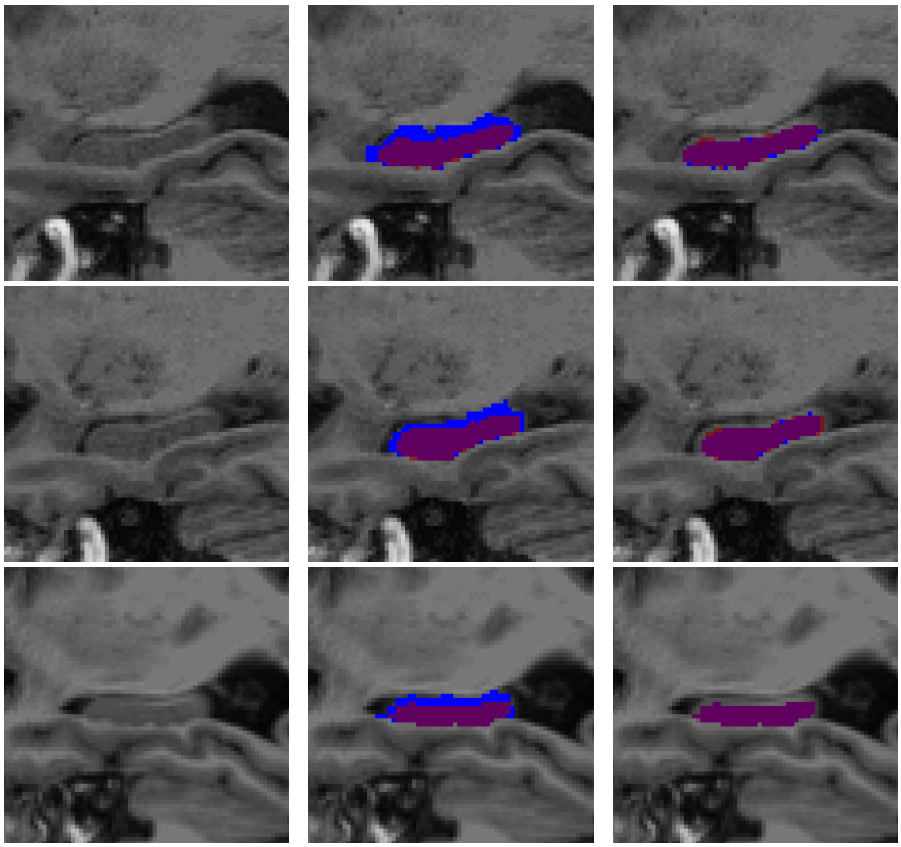

Figure 3: Examples of differences between the reference segmentations of the hippocampus and the automatic segmentations produced by FreeSurfer before and after applying the EC wrapper method. Red: reference segmentation; blue: automatic segmentation; purple: overlap region between automatic and reference segmentation. The EC method successfully corrects the spatial inconsistencies between the FreeSurfer results and the reference segmentations.

explains why the EC method was able to achieve a large improvement in segmentation accuracy.

To visualize and quantify the pattern of disagreement between FreeSurfer and reference segmentations across all subjects, we normalize the different hippocampus segmentations into a common coordinate space. Normalization is performed using a structure-specific shape-based normalization approach (Yushkevich et al., 2007). A geometrical model, known as the continuous medial representation (cm-rep), is fitted to each reference segmentation of the hippocampus, and the parameterization of this geometrical model is used to establish a one-to-one correspondence between the space inside and near the reference segmentation and a common reference space provided by a single template manual segmentation. Additional details on how correspondence maps are established using the cm-rep parameterization are given in the Appendix. Us-

ing these correspondence maps, we transfer both the reference segmentations and the FreeSurfer segmentations from subject space into the template space. We emphasize that since the same mapping is applied to both reference and automatic segmentations, the differences between these segmentations are maintained by the mapping. Averaging over all subjects, we compute the spatial label distribution in the template space for the reference segmentations and the FreeSurfer segmentations. These distributions are shown in Fig. 4. Note that a cm-rep model fitted to a reference segmentation does not overlap it perfectly. Hence, the mean spatial label distribution of the reference segmentations in the template space is not a binary image.

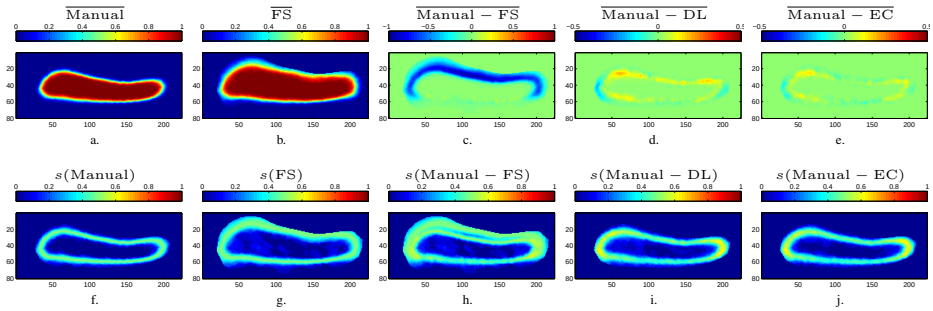

Figure 4: The spatial patterns of disagreement between the automatic segmentations of the hippocampus in the FreeSurfer experiment and the corresponding reference segmentations, plotted after normalization to a common reference space. All plots show a sagittal cross-section of the 3D reference space. (a). The mean of the normalized reference segmentations. (b). The mean of the FreeSurfer segmentations, mapped into the reference space using the same transformations as the corresponding reference segmentations. (c). Mean signed difference between FreeSurfer and reference segmentations. *FreeSurfer over-segmented the hippocampus at the superior, anterior and posterior boundaries.* (d). Mean signed difference between DL results and reference segmentations. (e). Mean signed difference between EC results and reference segmentations. *Both DL and EC methods correct the over-segmentation produced by FreeSurfer, the latter doing so more effectively.* (f-j). Standard deviation of the normalized reference segmentations, normalized automatic segmentations, and their signed differences.

The plot of mean signed difference between the normalized automatic and reference segmentations in Fig. 4(c) reveals a consistent pattern of disagreement in the anterior, posterior and superior regions of the hippocampus for the host method. Fig. 4(d) and 4(e) show that after applying the learning-based correction algorithms, this pattern of disagreement is reduced dramatically, with the EC method producing the greater reduction in disagreement than the DL method. Interestingly, neither EC nor DL completely eliminated consistent disagreement with reference segmentations, with both methods exhibiting a similar pattern of disagreement (slight under-segmentation at the anterior and posterior boundaries, and slight over-segmentation along the inferior and superior boundaries). Note that the learning algorithm and the host segmentation method produced different consistent errors. For example, the DL algorithm produces some consistent under segmentation around the anterior and posterior regions of the hippocampus, shown in yellow and red colors, however such consistent errors do not appear in the results produced by FreeSurfer and are significantly reduced in the results produced by error correction. This result suggests that including the host segmentation method's output in learning helps reduce the consistent errors only imposed by the

pure learning algorithm. Hence, our error correction method outperformed the direct learning algorithm.

A consistent pattern of difference between the automatic and reference segmentations can also be seen by examining the volumes of the segmentations. Table 4 shows the average hippocampal volume produced by FreeSurfer, which is almost double the average volume of the reference segmentations. Fig. 5 (left) plots the volume correlation between the reference segmentations and the FreeSurfer segmentations, revealing a substantial bias between the volume measurements, as well as the large variance in the difference between the measurements. Fig. 5 (right) plots the volume correlation after applying the EC method. Both the bias and the variance of the volume difference are dramatically reduced. After error correction, the volume differences between reference segmentation and automatic segmentation appear to be more similar to zero-mean random noises, as shown in Fig. 5.

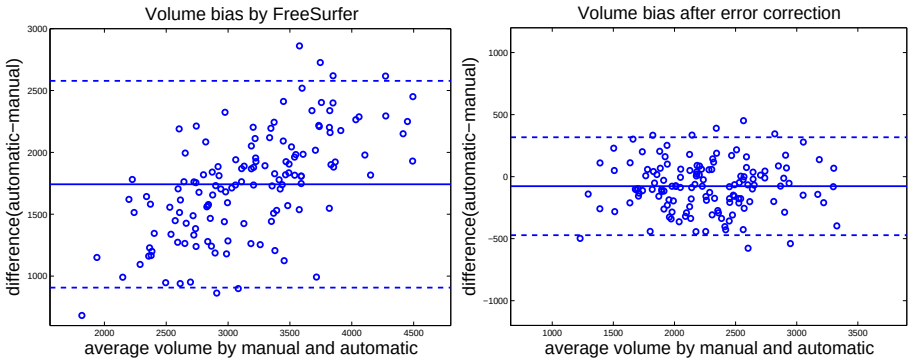

Figure 5: Bland-Altman plots comparing automatic volume (in mm^3) estimates produced by FreeSurfer to manual volume estimates. Each point corresponds to a hippocampal segmentation of one of the two hemispheres in one subject. The difference between automatic and manual estimates is plotted against their average. The solid horizontal line corresponds to the average difference, and the dashed lines are plotted at average ± 1.96 standard deviations of the difference

LEFT				RIGHT			
manual	FreeSurfer	DL	EC	manual	FreeSurfer	DL	EC
1952± 377	3349±599	1919±388	1904±384	1894±391	3412±598	1858±382	1843±385

Table 4: The average hippocampal volumes (mm^3) derived from the reference segmentations, the FreeSurfer segmentations, and the segmentations produced by the DL and EC wrapper methods.

The volume expansion bias observed in our experiment also has been reported in the literature, as summarized in Table 5. This may be due to the fact that FreeSurfer was trained on a manual hippocampal segmentation protocol different from ours and from those used by the other authors reporting the volume expansion bias. Besides the contribution from employing different manual segmentation protocols, Table 5 also reveals another potential bias associated with imaging modalities. FreeSurfer tends

to produce smaller volume expansion and higher segmentation overlap with manual segmentations on 1.5 T MR images than on 3 T MR images.

methods	MRI Field Strength	relative volume difference	Dice overlap
(Fischl et al., 2002)	N/A	∼105%	N/A
(Khan et al., 2008)	1.5 T	N/A	0.70 to 0.85
(Cherbuin et al., 2009)	1.5 T	∼125%	N/A
(Morey et al., 2009)	1.5 T	∼120%	∼0.82
(Morra et al., 2009)	1.5 T	N/A	0.73
(Sanchez-Benavides et al., 2010)	1.5 T	∼103%	∼0.78
(Akhondi-Asl et al., 2010)	3 T	∼150%	0.63
(Pardoe et al., 2009)	3 T	∼140%	∼0.7
ours	3 T	∼170%	∼0.66
ours after EC	3 T	∼97%	∼0.86

Table 5: Hippocampal segmentations produced by FreeSurfer reported in the recent literature. The results are summarized in terms of relative volume difference, i.e. the volume ratio between automatic and manual segmentations, and Dice overlap compared to manual segmentations. The volume ratios are estimated based on the volumes of automatic and manual segmentations reported in the corresponding work.

As discussed in the Introduction, all these factors are natural sources of consistent errors that a segmentation tool may produce when applied to a large variety of applications. One way to correct these errors would be to retrain and retune FreeSurfer on our data set. Our error correction method gives a simple alternative approach to adapt FreeSurfer to our data.

Effect of working ROI size. To investigate the influence of the working ROI size on the learning algorithms, we repeat the above experiment with working ROIs of different size. Two additional working ROIs were obtained from the FreeSurfer segmentations by applying dilations of two and three voxels. Since the working ROIs obtained from one-voxel dilation already cover most of the manually labeled hippocampi, these two larger working ROI definitions include more background voxels into consideration. Table 6 shows the percentage of background voxels and hippocampal voxels covered by the three working ROI definitions.

Fig. 6 shows the average Dice overlaps over 10 cross-validation experiments using these three different working ROIs. For the EC algorithm, using larger working ROIs results in almost identical but slightly worse results. By contrast, DL performance is

Dilation Radius	working ROI size (voxels)	% Hippocampal voxels in ROI	% ROI voxels in hippocampus
1-voxel	5978	99.67	32.19
2-voxel	8904	99.97	21.68
3-voxel	12234	99.99	15.78

Table 6: The ability of working ROIs generated with different radii of dilation to cover the reference hippocampal segmentations.

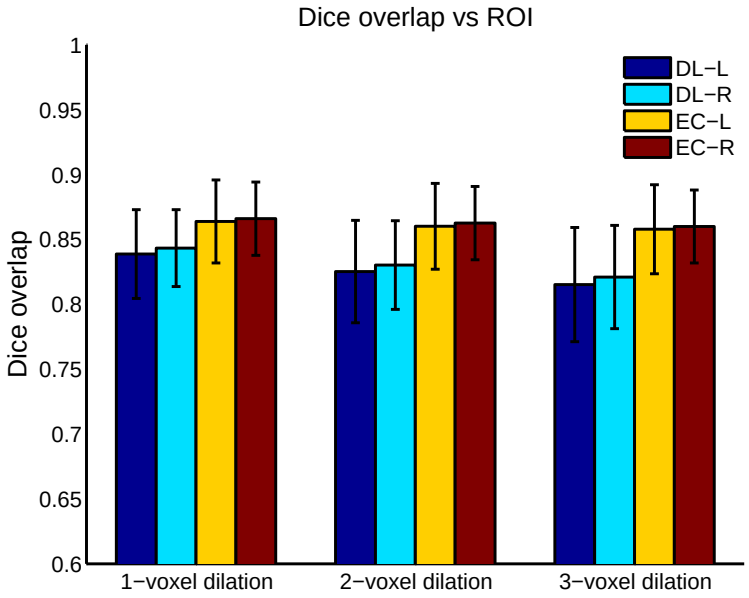

Figure 6: Dice overlaps of hippocampal segmentation when different dilations are used to generate the working ROIs for direct learning and error correction. Results for left and right sides are displayed separately.

more significantly reduced when larger working ROIs are used. This result implies that the location of the initial segmentation produced by the host method is informative for the segmentation problem. Ignoring this information by including more irrelevant background voxels indeed complicates the learning problem. Hence, in the remaining experiments, we restrict ourselves to use small working ROIs that have good coverage of the manually labeled foreground in the training data.

Effect of training set size. The experiments above use reference segmentations from 70 subjects for training. Such a large training set may be impractical for real-world applications. To investigate the effect of training set size on the error correction performance, we performed experiments using various numbers of training subjects (1-5, 10, and 20). To facilitate the comparison with the earlier results, the 10 cross-validation experiments above were repeated with the same partitioning of the subjects into training and test sets. However, in each experiment, only a subset of the full training set was used to train classifiers.

Fig. 7 shows the error correction performance with respect to the number of training data. Even using a single image for training, our method achieves a significant improvement over the host method. As more subjects are added to the training set, the segmentation performance increases, although with diminishing returns.

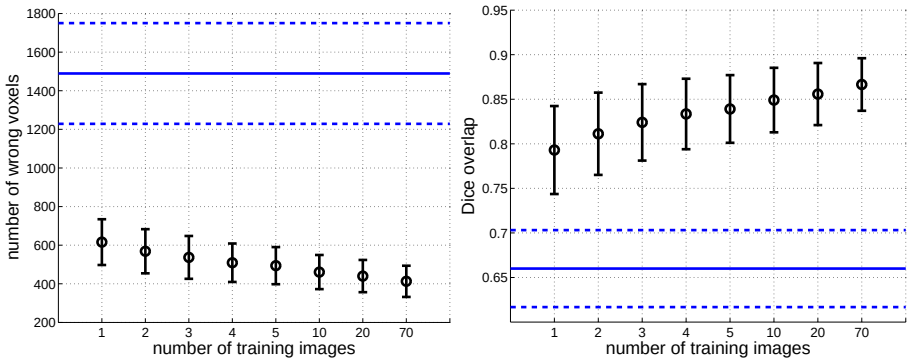

Figure 7: Effect of training set size on EC performance. Left: the average number of mislabeled voxels after error correction vs. the size of the training set (error bars at ± 1 s.d.). The average number of mislabeled voxels before error correction, ± 1 s.d., is shown in blue. Right: Dice overlap between EC results and reference segmentations vs. training set size, with the blue line showing Dice overlap without error correction.

3.2. Multi-Atlas Segmentation with Label Fusion

Fig. 8 shows the results of the parameter selection experiment for one of the cross-validation experiments. The optimal parameters for this experiment, found using leave-one-out analysis among the atlases, were $r = 2$ and $\sigma = 0.05$. Optimal parameters found for the remaining 9 cross-validation experiments were similar, with $r \in [2, 3]$ and $\sigma \in [0.05, 0.1]$.

On average, the multi-atlas approach produced 372 mislabeled voxels for each hippocampus. The working ROI for EC and DL training was obtained by a single-voxel dilation of the host segmentation results. On average, this ROI covered 98.7% of the manually labeled hippocampal voxels. The EC method produced 13.7% fewer errors than the multi-atlas method alone (321 mislabeled voxels). By contrast, DL produced worse results with 435 mislabeled voxels. Fig. 10 shows the spatial patterns of disagreement between automatic and reference segmentation before and after applying EC and DL. Consistent under-segmentation near the head and tail of the hippocampus is reduced substantially by the EC method, while the DL method introduces a new pattern of over-segmentation along the hippocampus boundary. Fig. 9 shows the volume differences after applying the error correction. The negative volume bias is significantly reduced. Table 7 shows the average Dice overlap for each of the cross-validation experiments. Overall, EC improves the Dice overlap from 0.88 to 0.90. The improvement is statistically significant, with $p < 0.00001$ on the paired t-test.

Fig. 11 shows the performance of the multi-atlas approach and the error correction method with respect to the number of atlases used. The numbers of atlases tested are 2, 3, 4, 5, 10, and 20. The experiment with one atlas is not included because using one atlas cannot produce training data for the error correction algorithm. Note that using fewer atlases, the multi-atlas label fusion technique produced less accurate segmentations. As the number of atlases increased, the segmentation accuracy also increased but with reduced increasing rates. This observation is consistent with previous studies, such as (Heckemann et al., 2006), that quantified the relationship between the number

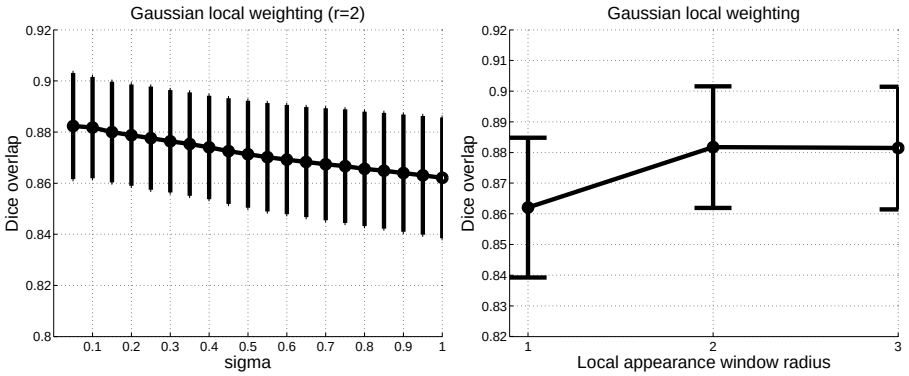

Figure 8: Performance of the image similarity based local weighting technique. The left plot shows the performance w.r.t. the Gaussian weight function when the local appearance window has $r = 2$. The right plot shows the best results achieved at different local appearance windows.

of atlases used for label fusion and segmentation accuracy. On the other hand, using fewer atlases also provided fewer training data for the EC algorithm. In this test, similar improvements were produced by the error correction technique when different numbers of atlases were used.

3.3. Brain Extraction

Since BET segmentations were well aligned with the manual segmentations, and most segmentation errors were cases of background mislabeled as brain tissue, we defined the working ROI by performing a one-voxel dilation of the brain segmentation produced by BET. On average, this ROI covered 99.3% of the foreground voxels in the manual segmentation.

An example of the segmentation improvement produced by the EC method is shown in Fig. 12. On average, each brain contains 9.7×10^5 voxels. BET produced 1.1×10^5 mislabeled voxels. EC produced 29% fewer errors (8.0×10^4 mislabeled voxels); by contrast, DL produced worse segmentations with 9.1×10^4 mislabeled voxels. Table 8 shows the results in terms of average Dice overlap for each of the 10 cross-validation experiments. The improvement achieved by EC over the host method is significant, with $p < 0.00001$ on the paired t-test.

Fig. 13 shows the error correction performance with respect to the number of training data. Again, the pattern noted in hippocampal segmentation experiments was observed. Although BET already produces good brain extraction results, the wrapper algorithm still could make significant improvements with only one training image.

3.4. Brain Tissue Segmentation

The manual segmentation protocol in IBSR presents certain challenges for evaluating segmentation methods. The protocol merges extraventricular CSF into the gray matter label (see Fig. 14). Consequently, a voxel labeled as CSF by FAST should be considered correctly labeled if that voxel has the grey matter label in the IBSR manual segmentation. To allow quantitative evaluation, we merge CSF into the gray matter

	LEFT			RIGHT		
Exp.	initial(Dice)	DL(Dice)	EC(Dice)	initial(Dice)	DL(Dice)	EC(Dice)
1($\sigma = 0.05, r = 2$)	0.878±0.036	0.858±0.025	**0.896**±0.030	0.874±0.017	0.859±0.022	**0.895**±0.020
2($\sigma = 0.05, r = 3$)	0.877±0.029	0.866±0.026	**0.895**±0.025	0.863±0.040	0.869±0.028	**0.893**±0.032
3($\sigma = 0.1, r = 2$)	0.885±0.026	0.866±0.023	**0.904**±0.020	0.869±0.039	0.861±0.034	**0.892**±0.036
4($\sigma = 0.05, r = 3$)	0.881±0.026	0.867±0.026	**0.902**±0.025	0.866±0.043	0.868±0.029	**0.892**±0.038
5($\sigma = 0.05, r = 3$)	0.886±0.020	0.870±0.024	**0.906**±0.017	0.877±0.026	0.872±0.025	**0.901**±0.023
6($\sigma = 0.1, r = 2$)	0.889±0.030	0.866±0.034	**0.904**±0.029	0.882±0.025	0.869±0.025	**0.904**±0.023
7($\sigma = 0.1, r = 3$)	0.891±0.018	0.868±0.024	**0.904**±0.019	0.869±0.022	0.873±0.026	**0.897**±0.021
8($\sigma = 0.1, r = 3$)	0.883±0.027	0.879±0.017	**0.908**±0.019	0.873±0.031	0.870±0.023	**0.895**±0.027
9($\sigma = 0.1, r = 2$)	0.896±0.024	0.872±0.024	**0.910**±0.020	0.882±0.023	0.870±0.024	**0.900**±0.021
10($\sigma = 0.05, r = 2$)	0.890±0.016	0.874±0.019	**0.908**±0.013	0.872±0.024	0.876±0.016	**0.903**±0.021

Table 7: Results of automatic hippocampus segmentation using the multi-atlas label fusion method and the wrapper methods. Each row gives the average Dice overlap with reference segmentation for one cross-validation experiment. The bold font highlights the best results.

Exp.	initial(Dice)	DL(Dice)	EC(Dice)
1	0.941±0.032	0.954±0.033	**0.959**±0.031
2	0.956±0.010	0.967±0.007	**0.972**±0.007
3	0.935±0.036	0.943±0.034	**0.948**±0.036
4	0.952±0.018	0.968±0.011	**0.973**±0.010
5	0.955±0.017	0.967±0.009	**0.971**±0.008
6	0.946±0.032	0.960±0.030	**0.964**±0.029
7	0.950±0.024	0.967±0.025	**0.970**±0.024
8	0.945±0.030	0.957±0.028	**0.961**±0.027
9	0.951±0.016	0.959±0.012	**0.963**±0.009
10	0.944±0.031	0.957±0.028	**0.962**±0.028

Table 8: Results of automatic brain extraction using BET and the wrapper methods. Each row gives average Dice overlap with manual segmentation for one cross-validation experiment. The bold font highlights the best results.

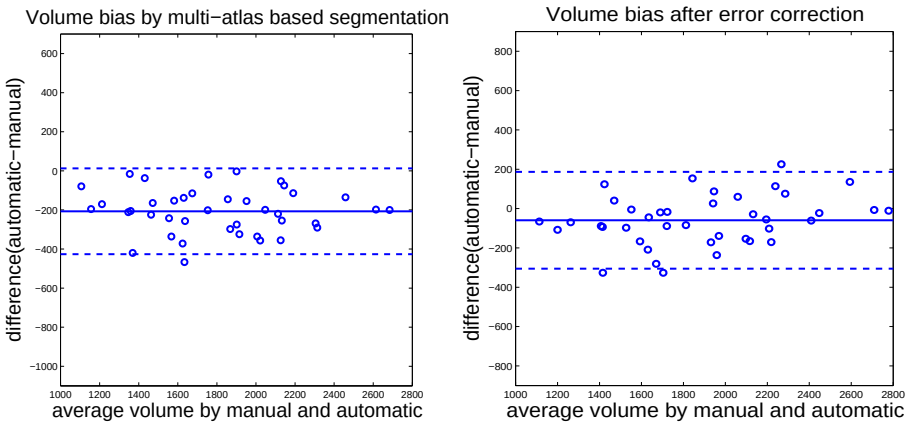

Figure 9: Bland-Altman plots comparing automatic volume estimates produced by multi-atlas based segmentation to manual volume estimates. Each point corresponds to a hippocampal segmentation of one of the two hemispheres in one subject. The difference between automatic and manual estimates is plotted against their average. The solid horizontal line corresponds to the average difference, and the dashed lines are plotted at average ± 1.96 standard deviations of the difference

label for both manual and automatic segmentation, and report overlaps for the white matter and merged gray matter labels. Note that our wrapper methods could be directly applied with any manual segmentation protocol. Merging CSF into gray matter is done purely to avoid an unfair comparison with FAST.

Fig. 14 shows segmentation examples produced by the FAST algorithm and the two wrapper algorithms. The average volume of the brain ROI was 9.7×10^5 voxels (recall from the Methods section that the working ROI for this experiment is the manual brain mask from IBSR). On average, FAST produced 8.9×10^4 mislabeled voxels. For explicit EC, the error detection step produced the precision of 92% (i.e., the fraction of voxels marked mislabeled that were actually mislabeled) with the recall of 84% (i.e., the fraction of mislabeled voxels that were detected). The error correction step correctly assigned new labels to 91% of the detected mislabeled voxels. Overall, explicit EC produced 21% fewer errors than FAST (7.0×10^4 mislabeled voxels). Implicit EC produced 17% fewer errors than FAST (7.4×10^4 mislabeled voxels). Since the gray matter and white matter have good appearance contrast, error detection achieved high accuracy. As a result, explicit EC outperforms implicit EC, despite having a lower computational cost. By contrast, DL produced worse segmentations with 8.1×10^4 mislabeled voxels. Again, the improvements achieved by implicit EC and explicit EC are significant, with $p < 0.00001$ on the paired t-test. Table 9 shows the results in terms of Dice overlap for each of the 10 cross-validation experiments. Like in the previous experiments, the error correction methods outperformed DL and the host segmentation method.

Fig. 15 shows the error correction performance with respect to the number of training data for both EC algorithms. When only one or two training subjects were used, implicit EC produced slightly better performance than explicit EC. These results suggest that due to the limited number of training subjects, error detection could not be

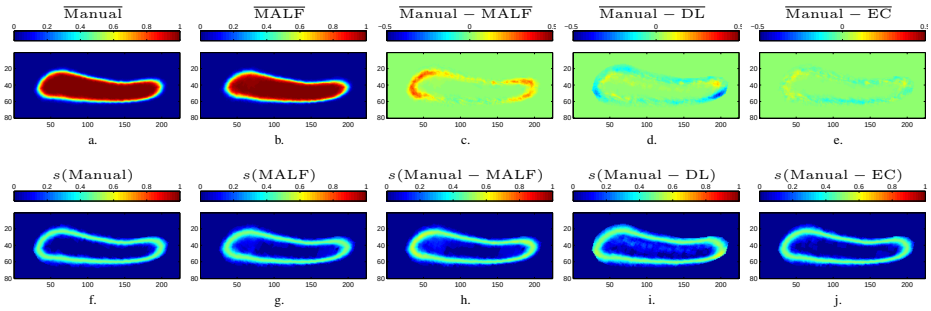

Figure 10: The spatial patterns of disagreement between the automatic segmentations of the hippocampus in the multi-atlas label fusion experiment (MALF) and the corresponding reference segmentations, plotted after normalization to a common reference space. See caption to Fig. 4 for details.

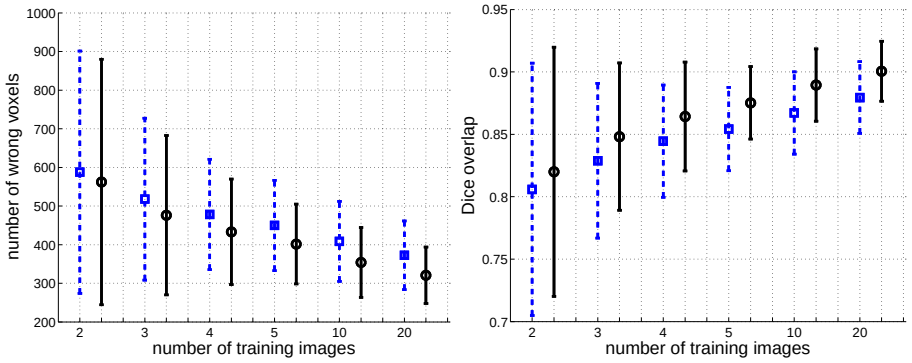

Figure 11: The figure on the left plots the average number of mislabeled voxels after error correction vs. the size of the training set (error bars at ± 1 s.d.). The average number of mislabeled voxels before error correction, ± 1 s.d., is shown in blue. The figure on the right similarly plots Dice overlap between error correction results and reference segmentations vs. training set size, with the blue line showing Dice overlap without error correction.

reliably done. However, when more than two training subjects were used, explicit EC produced slightly better results than implicit EC. Overall, significant improvements were achieved by both implicit and explicit EC when two or more training subjects were used, and using more training data consistently resulted in greater improvement.

4. Discussion

Across all four applications considered above, both EC algorithms achieved a significant improvement in accuracy relative to the manual segmentations. The number of mislabeled voxels was reduced by 72% for FreeSurfer hippocampus segmentation, 14% for multi-atlas hippocampus segmentation, 29% for BET brain extraction, and 21% for FAST brain tissue segmentation. In each experiment, both EC methods outperformed the DL method, demonstrating that including the results of the host method

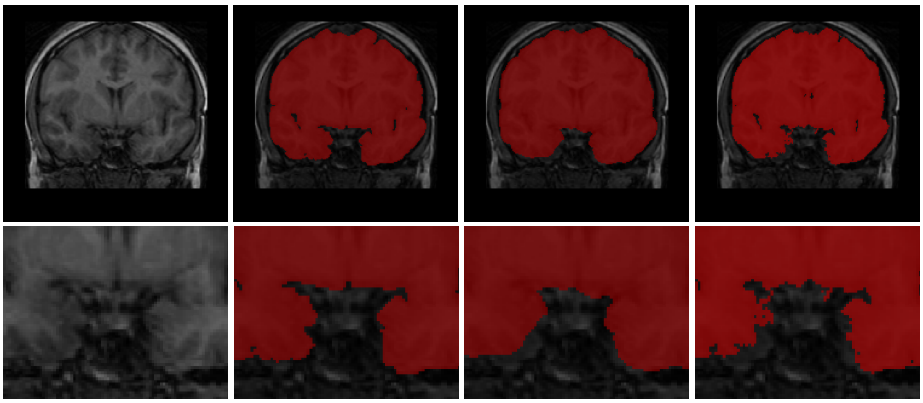

Figure 12: Brain extraction on T1-weighted MR images. Left to right: original image, manual brain extraction, initial brain extraction produced by BET, final segmentation produced by the EC method. The second row gives a zoomed in view.

segmentation as contextual features improves classifier performance. However, as Figures 4(d,e) and 10(d,e) show, neither DL nor EC are capable of completely eliminating consistent differences between automatic and reference segmentations. This suggests that some aspects of these consistent differences are too complex for AdaBoost to learn, at least using the features employed in this paper. In all four experiments, improvements were achieved even when very few images were used for training. However, larger training sets consistently led to improved performance. Overall, these results suggest that the EC method is capable of consistently improving segmentation performance across a broad range of medical image segmentation problems.

4.1. Comparison to the State of the Art Segmentation Methods

Segmentation performance produced by the EC method compares favorably with the state of the art in published work. Before making such a comparison, we echo the point made by Collins and Pruessner (2010) that direct comparisons of Dice overlaps and other quantitative segmentation quality measures across publications are difficult and not always fair, as these measures depend not only on the ability of the automatic method to mimic the human expert, but also on the underlying segmentation protocol, the imaging protocol, and the patient population. Nevertheless, the comparisons carried out below indicate the highly competitive performance achieved by combining host methods with error correction.

Segmentation of the hippocampus. Due to the central role of the hippocampus in memory encoding and its vulnerability to neurodegenerative diseases, there has been intense interest in MRI-based hippocampal morphometry. Many automatic approaches for hippocampus segmentation have been proposed, e.g. (Carmichael et al., 2005, Hammers et al., 2007, Powell et al., 2008, Morra et al., 2008, van der Lijn et al., 2008, Morra et al., 2009, Chupin et al., 2009, Pluta et al., 2009, Wolz et al., 2009, Collins and

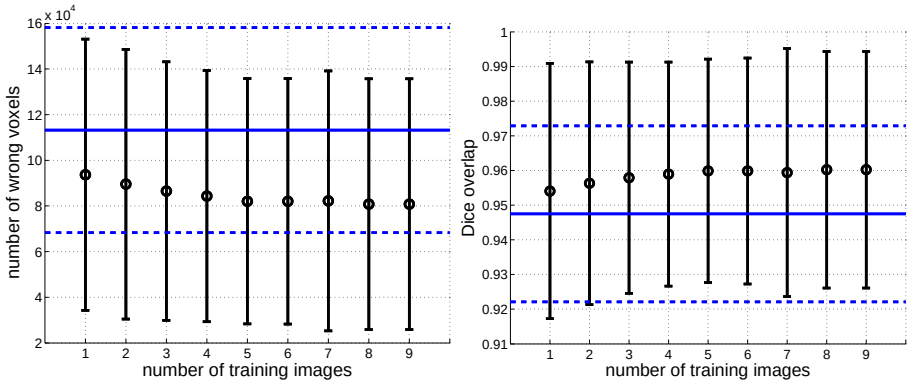

Figure 13: Effect of training set size on EC performance in the brain extraction experiment. Left: the average number of mislabeled voxels after error correction vs. the size of the training set (error bars at ± 1 s.d.). The average number of mislabeled voxels before error correction, ± 1 s.d., is shown in blue. Right: average Dice overlap between EC results and manual segmentations vs. training set size, with the blue line showing Dice overlap for BET without error correction.

	Gray Matter				White Matter			
Exp.	initial	DL	IEC	EEC	initial	DL	IEC	EEC
1	0.939±0.010	0.949±0.006	0.951±0.009	**0.954**±0.008	0.882±0.022	0.898±0.010	0.901±0.013	**0.907**±0.012
2	0.935±0.005	0.944±0.008	0.946±0.006	**0.949**±0.006	0.874±0.019	0.890±0.022	0.896±0.019	**0.902**±0.019
3	0.938±0.006	0.933±0.022	0.941±0.016	**0.941**±0.021	0.882±0.011	0.879±0.026	0.892±0.017	**0.894**±0.022
4	0.936±0.007	0.945±0.008	0.948±0.007	**0.951**±0.007	0.880±0.014	0.892±0.011	0.899±0.010	**0.905**±0.010
5	0.937±0.010	0.946±0.008	0.950±0.006	**0.953**±0.006	0.878±0.022	0.891±0.017	0.901±0.013	**0.907**±0.013
6	0.939±0.008	0.947±0.007	0.950±0.006	**0.954**±0.006	0.886±0.016	0.898±0.012	0.904±0.009	**0.910**±0.008
7	0.937±0.007	0.941±0.023	**0.949**±0.008	0.949±0.015	0.879±0.022	0.893±0.020	0.903±0.016	**0.906**±0.018
8	0.938±0.010	0.946±0.008	0.951±0.007	**0.954**±0.006	0.879±0.021	0.892±0.019	0.902±0.015	**0.909**±0.014
9	0.939±0.009	0.943±0.012	0.949±0.009	**0.951**±0.010	0.880±0.021	0.884±0.029	0.898±0.022	**0.902**±0.024
10	0.934±0.006	0.947±0.008	0.948±0.005	**0.951**±0.005	0.874±0.019	0.893±0.019	0.898±0.015	**0.904**±0.015

Table 9: Results of automatic brain extraction using FAST and the wrapper methods. Each row gives average Dice overlap with manual segmentation for one cross-validation experiment. The bold font highlights the best results.

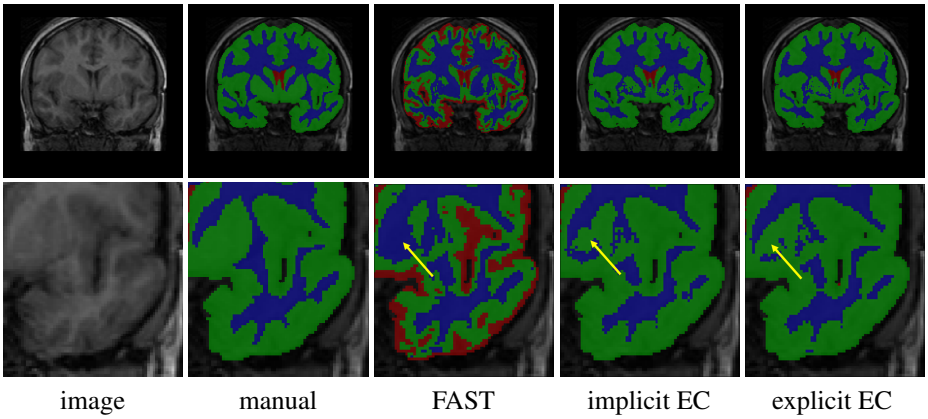

Figure 14: Brain tissue segmentation in T1-weighted MR images. Left to right: original image; manual segmentation consisting of white matter and gray matter and ventricle labels; three-tissue segmentation produced by the host method (FSL FAST); segmentation after correction by implicit EC; segmentation after correction by explicit EC. The second row displays a zoomed in view of the left temporal lobe. The arrow points to one significant correction made by the two EC methods. Note that the sulcal CSF in the FAST result is merged with the gray matter label for consistency with the manual segmentation (see text for details).

Pruessner, 2010, Leung et al., 2010). Table 10 summarizes the results of automatic hippocampus segmentation from recent publications. Most results are reported in terms of average Dice overlap, but a few are reported in terms of the average Jaccard index ($JI(A,B) = |A \cap B|/|A \cup B|$). A trend revealed in the table is that most of the best hippocampal segmentation approaches use multi-atlas label fusion. Furthermore, automatic hippocampal segmentation tends to reach better consistency with manual segmentations in healthy subjects than in subjects with neurodegenerative diseases.

Our experiment with FreeSurfer demonstrates how the error correction scheme can adapt a general segmentation tool to a different manual segmentation protocol and a different imaging modality. On the 3T MR images used in our experiment, FreeSurfer produced segmentations that overlapped poorly with our reference segmentations (the average Dice overlap was 0.660). However, the EC approach successfully adapted FreeSurfer segmentations better to match those created by the reference segmentation protocol. Dice overlaps with the reference segmentation after error correction (0.865 on average) are competitive with many of hippocampal segmentation results for normal controls and MCI patients published in the last few years.

However, our most competitive hippocampus segmentation results were achieved by pairing the EC method with multi-atlas label fusion. According to Table 10, the best published results for hippocampus segmentation to date have been produced by Collins and Pruessner (2010) and Leung et al. (2010). Both papers use multi-atlas segmentation. Collins and Pruessner (2010) evaluate segmentation performance using a leave-one-out strategy on 80 normal controls. Leung et al. (2010) use a template library of 55 atlases; however, for each atlas, both the original image and its flipped mirror image are used as atlases. Hence, Leung et al. (2010) effectively use 110 atlases for label fusion. Our multi-atlas approach uses only 20 atlases and, without error correction, produces results that are comparable to the state of the art for normal controls and are slightly

Methods and description	Dice	JI	Tested Cohort
(Heckemann et al., 2006): multi-atlas based segmentation	0.82		30 normal controls
(Hammers et al., 2007): multi-atlas based segmentation	0.76(sclerotic side) 0.83(contralateral side)		9 patients with unilateral hippocampal sclerosis
(Powell et al., 2008): machine learning based classification	0.85		15 subjects (with no population description)
(Barnes et al., 2008): multi-atlas based segmentation	0.87 0.86		19 normal controls, 36 AD patients
(Khan et al., 2008): single-atlas based segmentation with initialization by FreeSurfer	0.86		4 normal controls
(van der Lijn et al., 2008): multi-atlas + graph cuts	0.858		20 elderly subjects covering 7 population variation in 7 hippocampus size
(Morra et al., 2009): machine learning based classification	0.835 0.802		20 normal controls 20 AD patients
(Wolz et al., 2009): multi-atlas + graph cuts	0.860		20 normal controls + 20 MCI patients + 20 AD patients
(Chupin et al., 2009): landmark-guided single-atlas based segmentation, followed by a registration error detection and correction procedure	0.87 0.85 0.84		16 young normal controls 8 normal controls 8 normal controls + 8 with known hippocampal sclerosis + 7 with normal hippocampal volumes
(Collins and Pruessner, 2010): multi-atlas based segmentation	0.887		80 young normal controls
(Leung et al., 2010): multi-atlas based segmentation		0.80 0.81	10 normal controls 10 MCI patients
Multi-atlas based segmentation	0.887 0.872	0.798 0.774	57 normal controls 82 MCI patients
Multi-atlas+error correction	0.908 0.893	0.833 0.808	57 normal controls 82 MCI patients
FreeSurfer: single-atlas based segmentation	0.673 0.651	0.508 0.485	57 normal controls 82 MCI patients
FreeSurfer+error correction	0.877 0.859	0.782 0.754	57 normal controls 82 MCI patients

Table 10: Hippocampal segmentation performance reported in the recent literature compared to the results obtained by the EC wrapper method with FreeSurfer and multi-atlas label fusion as the host methods. The results are given in terms of Dice overlap ($Dice(A,B) = \frac{2|A \cap B|}{|A|+|B|}$) and Jaccard index ($JI(A,B) = |A \cap B|/|A \cup B|$). Note that Morra et al. (2009) report results by mixing controls and AD patients. They also report the performance for each diagnostic group relative to the mixed results. The results reported here for (Morra et al., 2009) are estimated from their reported results.

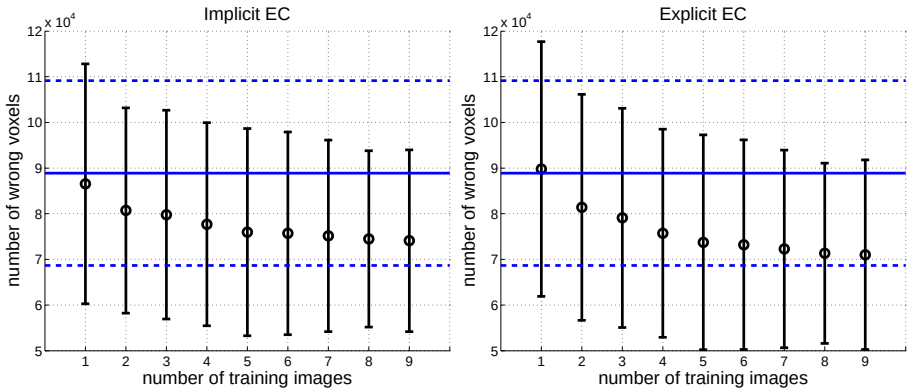

Figure 15: The average number of voxels mislabeled by the implicit and explicit variants of the EC method vs. the size of the training set (error bars at ± 1 s.d.). The average number of voxels mislabeled by the host method, ± 1 s.d., is shown in blue.

worse than the state of the art for MCI patients. With error correction, the results for both groups improve by $\sim$2% Dice overlap, yielding the same accuracy as the state of the art for MCI patients and better accuracy than the state of the art for normal controls. Thus, in so far as Dice overlaps across different methods can be compared, our method produces segmentation results as good or better than the state of the art, while requiring substantially fewer manually labeled images for training. However, unlike a number of other methods, our evaluation did not include AD subjects. Hence, it remains unclear how well multi-atlas segmentation with EC performs in the presence of severe hippocampal atrophy. Another potential limitation in our evaluation is that our gold standard segmentations were produced by manually editing automatic segmentations rather than manually segmenting each hippocampus from scratch. Hence, it is possible that our gold standard segmentations are influenced by the automatic segmentation results produced by (Pluta et al., 2009).

Brain extraction and tissue segmentation. Liu et al. (2009) evaluate a number of recent brain extraction methods, including (Shattuck and Leahy, 2002, Segonne et al., 2004, Zhuang et al., 2006), and their own method on the IBSR dataset. Their reported Dice overlap varies from 0.897 to 0.955. In our experiments, the BET algorithm produced competitive results with the average Dice overlap of 0.947. After EC, we produced Dice overlap of 0.964, which is almost 1% higher than the best results reported in (Liu et al., 2009).

For the brain tissue segmentation problem, Awate et al. (2006), Greenspan et al. (2006), Huang et al. (2009) report their methods' performance on the IBSR data set. However, the problem of inaccurate CSF labels was handled differently in each paper. Hence, only the white matter segmentation results are comparable between these papers. The average Dice overlaps in the white matter reported by these three papers are 0.887, 0.857, and 0.876 respectively. In our experiments, FAST produced competitive results with the average Dice overlap of 0.879. After error correction, we produced

0.905, which is almost 2% higher than the other methods.

Our improvements over these state of the art methods are at least comparable to the improvements introduced by these state of the art methods over their predecessors.

4.2. Relationship with Prior Work on Object Segmentation and Machine Learning

Our error correction method applies learning-based classification to perform segmentation, which is a commonly used technique in computer vision (Kumar and Hebert, 2003, Shotton et al., 2006, Tu and Bai, 2010) and medical image analysis (Tu et al., 2007, Morra et al., 2008). In particular, our work is closely related to (Morra et al., 2008, Tu and Bai, 2010), where instead of segmentation results produced by other segmentation methods, the results produced by earlier iterations of the learning algorithm itself are treated as high-level contextual features and are included in the learning process. By contrast, the main novel idea in our paper is to use machine learning as a *corrective technique* for segmentations produced by a given host method, rather than training classifiers from scratch to perform the segmentation problem. As a consequence, we report substantially better hippocampus segmentation results than Morra et al. (2009) for normal controls (0.908 vs. 0.835).

As in our approach, Chupin et al. (2009) also use an error correction procedure to improve the final results of their hippocampal segmentation application. However, their error correction approach was specially designed for their hippocampal segmentation problem. By contrast, our technique is general and can be easily applied to a wide range of segmentation problems.

4.3. Scenarios for Practical Application of Error Correction

The experiments in this paper illustrate two usage scenarios for the proposed method. In the first scenario, error correction is used to boost the performance of an existing automatic segmentation tool, provided example manual segmentations on a sample of the user's imaging data. As the FreeSurfer example illustrates, error correction is capable of adapting the host method to the particular imaging and segmentation protocols employed by the user, without the need to explicitly retrain the host method. This scenario is also illustrated by the experiments on whole brain segmentation. The drawback of this scenario is the need for the user to provide example manual segmentations. Thus, the benefit to users with small datasets may be limited. Such users may decide that if they were to embark on the path of manual segmentation, then they might as well segment all their images manually. However, for users with large datasets, the burden of segmenting the whole dataset manually is significantly larger than the cost of providing 10 or 20 training examples, while even a small improvement in segmentation accuracy may be of significant benefit. Additionally, many users with small or large datasets may be able to leverage existing segmentations from earlier studies, provided that they use similar imaging and anatomical protocols and cover similar subject populations. Lastly, the burden of manual segmentation may be reduced by manually editing the results of the host segmentation method, rather than generating manual segmentations from scratch.

For users with access to example manual segmentations, an alternative to error correction is to retrain or retune the host segmentation method. Based on the experiments

performed in this paper, it is not possible to predict which approach would lead to greater improvement in segmentation accuracy. However, retraining may be outside of the technical expertise for some users, or such an option might not be provided by the software implementation of the host method. By contrast, error correction can be performed relatively easily, and a reference open-source software implementation has been provided.

The second scenario, illustrated in the multi-atlas experiment, is for error correction to be incorporated into an image segmentation tool by the tool's developer. The multi-atlas experiment shows that even when there are no imaging or anatomical protocol differences between the data on which a host method is trained and the data to which it is applied, error correction can still improve performance significantly. Furthermore, when the host method is itself trained using example data, no additional example segmentations are needed for error correction, since training can be performed in a leave-one-out framework. Thus, error correction offers an opportunity to improve the performance of various existing segmentation tools at little additional cost to the developer, and virtually no cost to the user.

4.4. Limitations and Future Work

We have chosen a fairly straightforward approach to apply machine learning, which combines AdaBoost and simple intensity features. Other classifiers may perform better than AdaBoost, or they may be complimentary to it, leading to more accurate error correction. Likewise, including additional higher-order features may improve performance. For example, Haar filters offer information-rich and robust features that have been successfully applied to many other medical image applications (Tu et al., 2007, Morra et al., 2008). Using such features will be more appropriate for problems where there is a significant variation in the scale and orientation of the images.

In the current EC method, error correction is performed independently at each voxel. Imposing regularity conditions on the final segmentation after error correction, for instance using a Markov random field prior or a prior that enforces topological constraints, may further improve segmentation accuracy.

In our current approach, we use a single host method, a set of example segmentations from a single expert, and a single training set that combines all cohorts in a given study. A natural extension of the method is to combine results from multiple host methods, to provide the ability to handle segmentations by different experts, and to offer strategies for dealing with heterogeneous subject populations. The challenge in dealing with multiple host segmentations and multiple manual segmentations is that the number of patterns of systematic disagreement between manual and automatic segmentation grows quadratically. One way to address this problem is to derive consensus automatic and manual segmentations using a method such as STAPLE (Warfield et al., 2004), which would allow the current error correction method to be applied directly. However, this approach sacrifices much of the information contained in the original automatic and manual segmentations. Heterogeneous populations may be handled by training error correction classifiers separately for subjects with different diagnoses. Alternatively, diagnosis and demographic variables could be included as features for classifier training, allowing the classifier to learn the patterns of error that are common across the population.

5. Conclusions

We presented a simple but effective learning-based method for reducing the consistent errors that automatic tools make relative to manual segmentations. The main contribution over prior work on learning-based medical image segmentation was to include the results of the segmentation by a given host method as contextual features for classifier training. Our results, conducted in three different segmentation problems using four different host methods showed that the proposed approach consistently improves segmentation accuracy relative to manual segmentations, even when just a handful of training datasets are provided. Furthermore, by pairing our error correction method with well-established host segmentation methods, we obtained some of the best results published so far for hippocampus segmentation, brain extraction, and brain tissue classification. Anticipating that similar improvements can be obtained in other segmentation problems, we provided an open-source implementation of the error correction method and identified two usage scenarios, one targeting users of existing segmentation tools that seek to adapt these tools to their data, and the other targeting developers of training-based automatic segmentation algorithms.

Acknowledgements

The project described was supported by the Penn-Pfizer Alliance grant 10295, Award Number K25 AG027785 from the National Institute On Aging, and Award Number R21 NS061111 from the National Institure of Mental Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute On Aging, National Institute of Mental Health or the National Institutes of Health.

Data collection and sharing for this project was also funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc, F. Hoffman-La Roche, Schering-Plough, Synarc, Inc., and Wyeth, as well as non-profit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to ADNI are facilitated by the Foundation for the National Institutes of Health (http://www.fnih.org/). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129, K01 AG030514, and the Dana Foundation.

References

A. Akhondi-Asl, K. Jafari-Khouzanic, K. Elisevichd, and H. Soltanian-Zadeh. Hippocampal volumetry for lateralization of temporal lobe epilepsy: Automated versus manual methods. *NeuroImage*, In Press, 2010.

P. Aljabar, R. Heckemann, A. Hammers, J. Hajnal, and D. Rueckert. Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. *NeuroImage*, 46:726–739, 2009.

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

B. Avants, C. Epstein, M. Grossman, and J. Gee. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12:26–41, 2008.

S. Awate, T. Tasdizen, N. Foster, and R. Whitaker. Adaptive Markov modeling for mutual-information-based, unsupervised MRI brain-tissue classification. *Medical Image Analysis*, 10(5):726–739, 2006.

J. Barnes, J. Foster, R.G. Boyes, T. Pepple, E.K. Moore, J.M. Schott, C. Frost, R.I. Scahill, and N.C. Foxa. A comparison of methods for the automated calculation of volumes and atrophy rates in the hippocampus. *NeuroImage*, 40:1655–1671, 2008.

O. Carmichael, H. Aizenstein, S. Davis, J. Becker, P. Thompson, C. Meltzer, and Y. Liu. Atlas-based hippocampus segmentation in Alzheimer's Disease and mild cognitive impairment. *NeuroImage*, 27(4):979–990, 2005.

N. Cherbuin, K. Anstey, C. Reglade-Meslin, and P. Sachdev. In vivo hippocampal measurement and memory: A comparison of manual tracing and automated segmentation in a large community-based sample. *PLoS ONE*, 4(4):e5265, 2009.

M. Chupin, A. Hammers, R.S.N. Liu, O. Colliot, J. Burdett, E. Bardinet, J.S. Duncan, L. Garnero, and L. Lemieux. Automatic segmentation of the hippocampus and the amygdala driven by hybrid constraints: Method and validation. *NeuroImage*, 46: 749–761, 2009.

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

B. Fischl, D. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, A. Killiany, A. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen, and A. Dale. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33:341–355, 2002.

Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the* 2_{nd} *European Conf. on Computational Learning Theory*, pages 23–27, 1995.

H. Greenspan, A. Ruf, and J.Goldberger. Constrained gaussian mixture model framework for automatic segmentation of MR brain images. *IEEE Trans. on Medical Imaging*, 25(9):1233–1245, 2006.

A. Hammers, R. Heckemann, M. J. Koepp, J. S. Duncan, J. V. Hajnal, D. Rueckert, and P. Aljabard. Automatic detection and quantification of hippocampal atrophy on MRI in temporal lobe epilepsy: A proof-of-principle study. *NeuroImage*, 36:38–47, 2007.

D. Hasboun, M. Chantome, A. Zouaoui, M. Sahel, M. Deladoeuille, N. Sourour, M. Duymes, M. Baulac, C. Marsault, and D. Dormont. MR determination of hippocampal volume: Comparison of three methods. *Am J Neuroradiol*, 17:1091–1098, 1996.

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

A. Huang, R. Abugharbieh, and R. Tam. A hybrid geometricstatistical deformable model for automated 3-D segmentation in brain MRI. *IEEE Trans. on Biomedical Engineering*, 56(7):1838–1848, 2009.

A. Khan, L. Wang, and M. Beg. FreeSurfer-initiated fully-automated subcortical brain segmentation in MRI using Large Deformation Diffeomorphic Metric Mapping. *NeoroImage*, 41(3):735–746, 2008.

S. Kumar and M. Hebert. Discriminative random fields: A discriminative framework for contextual interaction in classification. In *ICCV*, 2003.

K. Leung, J. Barnes, G. Ridgway, J. Bartlett, M. Clarkson, K. Macdonald, N. Schuff, N. Fox, and S. Ourselin. Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's Disease. *NeuroImage*, 51:1345–1359, 2010.

J. Liu, Y. Chen, and L. Chen. Accurate and robust extraction of brain regions using a deformable model based on radial basis functions. *Journal of Neuroscience Methods*, 183:255–266, 2009.

R. Morey, C. Petty, Y. Xu, J. Hayes, H. Wagner II, D. Lewis, K. LaBar, M. Styner, and G. McCarthy. A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. *NeoroImage*, 45:855–866, 2009.

J. Morra, Z. Tu, L. Apostolova, A. Green, A. Toga, and P. Thompson. Automatic subcortical segmentation using a contextual model. In *Proceedings of the 11th international Conf. on Medical Image Computing and Compter-Aided Intervention*, pages 194–201, 2008.

J. Morra, Z. Tu, L. Apostolova, A. Green, A. Toga, and P. Thompson. Comparison of Adaboost and support vector machines for detecting Alzheimer's Disease through automated hippocampal segmentation. *IEEE Trans. on Medical Imaging*, 29(1):30–43, 2009.

H. Pardoe, G. Pell, D. Abbott, and G. Jackson. Hippocampal volume assessment in temporal lobe epilepsy: How good is automated segmentation? *Epilepsia*, 50(12): 2586–2592, 2009.

J. Pluta, B. Avants, S. Glynn, S. Awate, J. Gee, and J. Detre. Appearance and incomplete label matching for diffeomorphic template based hippocampus segmentation. *Hippocampus*, 19:565–571, 2009.

S. Powell, V. Magnotta, H. Johnson, V. Jammalamadaka, R. Pierson, and N. Andreasen. Registration and machine learning based automated segmentation of subcortical and cerebellar brain structures. *NeuroImage*, 39(1):238–247, 2008.

T. Rohlfing, R. Brandt, R. Menzel, D. B. Russakoff, and C. R. Maurer Jr. Quo vadis, atlas-based segmentation? *The Handbook of Medical Image AnalysisVolume III: Registration Models*, pages 435–486, 2005.

M. Sabuncu, B.T.T. Yeo, K. Van Leemput, B. Fischl, and P. Golland. A generative model for image segmentation based on label fusion. *IEEE Trans. on Medical Imaging*, In Press, 2010.

G. Sanchez-Benavides, B. Gomez-Anson, A. Sainz, Y. Vives andn M. Delfino, and J. Pena-Casanova. Manual validation of FreeSurfer's automated hippocampal segmentation in normal aging, mild cognitive impairment, and Alzheimer Disease subjects. *Psychiatry Research: Neuroimaging*, 181:219–225, 2010.

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

D. Shattuck and R. Leahy. BrainSuite: An automated cortical surface identification tool. *NeuroImage*, 6:129–142, 2002.

J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, 2006.

S. Smith. Fast robust automated brain extraction. *Human Brain Mapping*, 17(3):143–155, 2002.

S.M. Smith, M. Jenkinson, M.W. Woolrich, C.F. Beckmann, T.E.J. Behrens aand H. JohansenBerg, P.R. Bannister, M.D. Luca, I. Drobnjak, D.E. Flitney, R.K. Niazy, J. Saunders, J. Vickers, Y. Zhang, N.D. Stefano, J.M. Brady, and P.M. Matthews. Advances in functional and structural mr image analysis and implementation as fsl. *Neuroimage*, 23(Suppl 1):S208S219, 2004.

Z. Tu and X. Bai. Auto-context and its application to high-level vision tasks and 3D brain image segmentation. *IEEE Trans. on PAMI*, 32(10):1744–1757, 2010.

Z. Tu, S. Zheng, A. Yuille, A. Reiss, R. Dutton, A. Lee, A. Galaburda, I. Dinov, P. Thompson, and A. Toga. Automated extraction of the cortical sulci based on a supervised learning approach. *IEEE Trans. on Medical Imaging*, 26(4):541–552, 2007.

F. van der Lijn, T. den Heijer, M. M.B. Breteler, and W. J. Niessen. Hippocampus segmentation in MR images using atlas registration, voxel classification, and graph cuts. *NeuroImage*, 43:708–720, 2008.

P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages I–511–518, 2001.

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

S. Warfield, K. Zou, and W. Wells. Validation of image segmentation by estimating rater bias and variance. *Philosophical Transactions of the Royal Society*, 366(1874): 2361–2375, 2008.

R. Wolz, P. Aljabar, D. Rueckert, R. A. Heckemann, and A. Hammers. Segmentation of subcortical structures in brain MRI using graph-cuts and subject-specific a priori information. In *IEEE International Symposium on Biomedical Imaging*, pages 470–473, 2009.

P. Yushkevich, J. Piven, H. Hazlett, R. Smith, S. Ho, J. Gee, and G. Gerig. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *NeuroImage*, 31(3):1116–1128, 2006a.

P. Yushkevich, H. Zhang, and J. Gee. Continuous medial representation for anatomical structures. *IEEE Trans Med Imaging*, 25(2):1547–1564, 2006b.

P. Yushkevich, J. Detre, D. Mechanic-Hamilton, M. Fernndez-Seara, K. Tang, A. Hoang, M. Korczykowski, H. Zhang, and J. Gee. Hippocampus-specific fMRI group activation analysis using the continuous medial representation. *NeuroImage*, 35(4):1516–1530, 2007.

P. Yushkevich, B. Avants, S. Das, J. Pluta, M. Altinay, and C. Craige. Bias in estimation of hippocampal atrophy using deformation-based morphometry arises from asymmetric global normalization: An illustration in adni 3 t mri data. *NeuroImage*, 50(2):434–445, 2010.

Y. Zhang, M. Brady, and S. Smith. Segmentation of brain MR images through a hidden Markov random field model and the expectation maximization algorithm. *IEEE Trans. on Medical Imaging*, 20(1):45–57, 2001.

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

Appendix

Reliability of Manual Hippocampus Segmentation

To put our hippocampus segmentation results in the proper context, we report the reliability of our manual segmentation protocol. Table 11 summarizes the intra-rater and inter-rater reliability in terms of Dice overlap when the raters segmented the hippocampus from scratch. The reliability test was done on 10 randomly selected images.

intra-rater	inter-rater
0.900	0.865

Table 11: The average Dice overlap for intra-rater and inter-rater manual segmentation on 10 subjects when manual segmentation is done from scratch.

To efficiently obtain reference segmentations for the 139 ADNI images, the manual segmentations used in our experiment were obtained by a rater (MA) editing the automatic segmentation results produced by a semi-automatic method (Pluta et al., 2009). For this case, the intra-rater reliability test obtained an average 0.923 Dice overlap on 9 randomly selected images, slightly better than segmenting from scratch.

Shape-Based Normalization of the Hippocampus for Visualizing Patterns of Disagreement Between Automatic and Manual Segmentation

To normalize the reference and automatic segmentations of the hippocampus from multiple subjects to a common coordinate space, we employ the shape-based normalization approach from (Yushkevich et al., 2006b). A deformable cm-rep model is fitted to each binary reference hippocampus segmentation. The cm-rep model is a deformable model that explicitly specifies the medial axis of the hippocampus as a parametric surface, as well as the local thickness of the hippocampus as a parametric scalar field defined over the medial axis. The boundary of the hippocampus is derived from the medial axis and thickness scalar field analytically. This model is fitted to binary segmentations of the hippocampus by maximizing the overlap between the region enclosed by the model's boundary and the binary segmentation. The model imposes a 3D coordinate system on the interior of the hippocampus. The medial axis of the model is parameterized by a pair of variables μ_1 and μ_2, which denote two axes of the cm-rep coordinate system. For every location on the medial manifold, two line segments, called spokes, emanate and reach the boundary of the model. These line segments are orthogonal to the boundary; they completely span the model's interior, and thus provide the third axis in the cm-rep coordinate system, denoted by the variable ξ. ξ describes the relative depth of a point on a model's interior. It varies from 0 at points on the medial axis to +1 and -1 at points where the two spokes reach the boundary. Therefore, any point within the hippocampal volume is represented by the vector (μ_1, μ_2, ξ). The 3D coordinate system establishes a one-to-one correspondence between the interiors of models fitted to different hippocampus binary segmentations. To extend the correspondence to the exterior of the fitted models, we allow ξ to take values beyond $\pm$ 1. Additional details on shape-based correspondence using the cm-rep model are in (Yushkevich et al., 2006b).