Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Neuroinformatics. Author manuscript; available in PMC 2018 May 23.

Published in final edited form as:

Neuroinformatics. 2013 October ; 11(4): 447–468. doi:10.1007/s12021-013-9190-5.

Deformable Templates Guided Discriminative Models for Robust 3D Brain MRI Segmentation

Cheng-Yi Liu,

Laboratory of Neuro Imaging Department of Neurology, UCLA School of Medicine 635 Charles E. Young Drive South, Suite 225, 90095, Los Angeles, CA, USA

Juan Eugenio Iglesias, and

Department of Radiology Massachusetts General Hospital 149 13th street, ste. 2301, 02129 Charlestown, MA, USA

Zhuowen Tu

Laboratory of Neuro Imaging Department of Neurology, UCLA School of Medicine 635 Charles E. Young Drive South, Suite 225, 90095, Los Angeles, CA, USA

for The Alzheimer's Disease Neuroimaging Initiative

Abstract

Automatically segmenting anatomical structures from 3D brain MRI images is an important task in neuroimaging. One major challenge is to design and learn effective image models accounting for the large variability in anatomy and data acquisition protocols. A deformable template is a type of generative model that attempts to explicitly match an input image with a template (atlas), and thus, they are robust against global intensity changes. On the other hand, discriminative models combine local image features to capture complex image patterns. In this paper, we propose a robust brain image segmentation algorithm that fuses together deformable templates and informative features. It takes advantage of the adaptation capability of the generative model and the classification power of the discriminative models. The proposed algorithm achieves both robustness and efficiency, and can be used to segment brain MRI images with large anatomical variations. We perform an extensive experimental study on four datasets of T1-weighted brain MRI data from different sources (1,082 MRI scans in total) and observe consistent improvement over the state-of-the-art systems.

Keywords

Brain image segmentation; Fusion; Deformable templates; Discriminative models; Generative models

Correspondence to: Zhuowen Tu.

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at:http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Electronic supplementary material The online version of this article (doi: 10.1007/s12021-013-9190-5) contains supplementary material, which is available to authorized users.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Introduction

In neuroimaging studies, brain MRI segmentation is often a critical preprocessing step. Automated segmentation enables morphometric analysis of cortical and subcortical structures in large datasets (Fischl et al. 2002), a scenario in which manual labeling is impractical. The automatically segmented regions can be used to extract informative characteristics of structures, such as volumes and shape. In the clinic, these features have the potential to be used to evaluate the condition of a subject. Moreover, the identified boundaries between cortical and subcortical structures can aid the planning of brain surgery (Wels et al. 2009). In neuroscience research, statistics derived from the segmentations of control and experimental groups can be used to identify structural differences between them. In the context of disease studies, such differences can lead to the identification of new pathological biomarkers. For instance, the atrophy and morphological change of hippocampus have been identified as important markers for Alzheimer's disease (Jack et al. 2008).

Several practical segmentation methods are widely used in neuroimaging studies, e.g. Caret (2001), FreeSurfer (2002), FSL (2002), ITK-SNAP (2006), SPM (2003), and the segmentation utilities in 3D Slicer (Pieper et al. 2006). However, robustness against variations in the input imaging data remains an open problem. The main difficulties are due to: (1) variations in image intensities due to differences in MRI acquisition (hardware, pulse sequence, imaging parameters); (2) anatomical variations within and across populations. Intensity normalization and image registration (Hou 2006; Klein and et al. 2009) can be used to standardize the images prior to segmentation, but only to some extent, since many of the variations are intrinsic.

In this paper, we aim to build a robust system that automatically segments T1-weighted brain MRI volumes into anatomical sub-cortical and cortical structures. We approach the 3D brain image segmentation problem from a statistical modeling perspective, combining a generative and a discriminative model with feature augmentation and adaptation. Generative and discriminative models were first explored and compared in the machine learning and computer vision literatures (Ng and Jordan 2002; Tu 2007). It has been shown that, while generative models outperform discriminative models when the size of the training dataset is small, the latter often have a better asymptotic behavior (Liang and Jordan 2008). Works that attempt to combine both types of model include (Jebara 2003; Raina et al. 2003; Lasserre et al. 2006; Holub et al. 2008), which show that integrating the two types of models can be beneficial.

Specifically, the goal of brain MRI segmentation is to separate the voxels of an input scan into a number of classes. In some studies, these classes correspond to the three basic tissue types in the brain: gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) (Wells et al. 1996; Pham and Prince 1999; Leemput et al. 2001; Shattuck et al. 2001; Wu and Chung 2005; Bazin and Pham 2007; Li and Fan 2008). Other works have attempted to produce labels at the level of brain structures (e.g., hippocampus, pallidum, putamen, etc.) (Fischl et al. 2002; Scherrer et al. 2007; Klauschen et al. 2009; Bazin and Pham 2009), which is a more difficult problem but yields a richer description of the data.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

To produce these labels, generative models typically rely on two components: a *prior* term that summarizes the statistical frequency and spatial distribution of labels and a *likelihood* term that models how these labels translate into intensities. Then, Bayesian inference can be used to estimate which labels (i.e., which segmentation) are the most likely given the observed image intensities. The prior term is usually in the form of a statistical atlas endowed with deformation model (Ashburner and Friston 2005; Leemput et al. 2001; Fischl et al. 2002). Other priors include the use of principal component analysis (PCA) to model whole shape variations (Pohl et al. 2006; Yang et al. 2004) or Markov random fields (MRF) to impose local shape constraints (Fischl et al. 2002; Woolrich and Behrens 2006; Scherrer et al. 2009; Caldairou et al. 2011). For the likelihood term, Gaussian distributions or mixtures thereof have been predominant in the literature, due to their ease of inference (Fischl et al. 2004; Yang et al. 2004; Pizer et al. 2003; Corso et al. 2008). Because the Gaussian parameters (means and variances) are estimated during the optimization, these algorithms are robust against changes in MRI contrast. Moreover, they can also explicitly model image artifacts such as the MRI bias field, making them robust against them as well (see for instance Leemput et al. (2001)).

Recently proposed multi-atlas methods such as Gouttard (2007), Wu and Chung (2008), Klein et al. (2009), Aljabar et al. (2007), Heckemann et al. (2006), Wolz et al. (2009), Bazin and Pham (2009), and Sabuncu et al. (2010) can also be seen as generative models. These methods are based on deforming a number of labeled templates to a test scan, and then fusing the deformed labels into a single, enhanced estimate of the segmentation. As explained in Sabuncu et al. (2010), these algorithms can be interpreted as generative models in which the intensity and label of each voxel are taken from one of the deformed templates as indexed by a latent, discrete field.

Discriminative models attempt to directly estimate the label of each voxel given the local appearance of the image around it. To do so, a number of features are computed for each voxel and fed to a classifier that attempts to infer the corresponding label. Popular choices of features include image intensities, gradients, textures and other derived local measures (Tu et al. 2008). Choices of classifier range from simple rule-based classifiers (Li et al. 1993) to more complicated frameworks such as support vector machines (SVM Lee et al. 2005; Lao et al. 2006; Akselrod-Ballin et al. 2006; Zhang et al. 2009), AdaBoost (Quddus et al. 2005; Morra et al. 2010) and the increasingly popular random forests (Breiman 2001; Yi et al. 2009; Geremia et al. 2010; Yaqub et al. 2011). These techniques show promising results when segmenting tissues, structures, and even tumors (Bauer et al. 2011; Li and Fan 2012) if the variations of the test data with respect to the training data are relatively small. Unfortunately, changes in MRI contrast due to differences in imaging hardware or acquisition protocol considerably reduce the performance of these methods, limiting their applicability to MRI scans that have not been acquired the same way as the training dataset.

Comparing generative and discriminative models, we observe that deformable templates (Felzenszwalb 2005; Pizer et al. 2003) guided by generative models can efficiently model anatomical structures, thanks to their flexibility and adaptability. However, their simplified assumptions on underlying image statistics (e.g., Gaussian distributions) limit their ability to deal with complex intensity patterns. On the other hand, such patterns can be efficiently

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

captured by discriminative models thanks to their capacity of fusing together a large number of features. However, as described above, they are sensitive to global intensity changes and have difficulties to include spatial information. In this paper, we propose combining the strengths of the two approaches by fusing deformable templates (generative) and informative features (discriminative). The presented approach is based on using the estimated segmentation and the parameters of the generative model to normalize the image intensities and extract robust, invariant local features. This creates an augmented feature space that can be used in a discriminative framework, effectively fusing the two types of model.

The rest of this paper is organized as follows. First, section “Further Related Work” surveys other attempts of combining generative and discriminative models in brain MRI segmentation. Section “Model Description” describes the proposed segmentation framework. Section “Learning and Using Fusion Models” describes how to train the model and use it to segment previously unseen test cases. A set of experiments and their corresponding results are described in section “Experiments”. Finally, section “Conclusion and Future Work” concludes the paper.

Further Related Work

Here we discuss other works in the literature that are similar to the proposed approach, highlighting the differences between them:

- Verma et al. (2008) use Bayesian and SVM models to classify the intra- and inter-patient tissue distributions. However, the two families of methods are separately used for the different sub-problems, rather than in an integrated fashion.
- In Tu et al. (2008), a discriminative classifier is used for appearance, and its classification results are regularized by generative models (PCA) capturing the shape prior. The two types of models are represented as separate modules and not jointly considered. In this paper, we utilize the same discriminative model, whereas the generative models are totally different. As shown in the experiments later on in this paper, the proposed method outperforms this algorithm, which is not robust across datasets due to the inability of the discriminative model to adapt to the new data.
- Wels et al. (2009) cast the subcortical segmentation as a posterior maximizing problem with shape parameters. The whole problem is decomposed into four sub-problems in sequence and each is solved by a discriminative model. Their segmentation then locates, rotates, scales, and refines the boundaries of structures in order, which performs a rough generative process. Nevertheless, the four underlying classifiers are purely discriminative so their weaknesses remain. If some stage fails, wrong information is propagated without any adaptation or correction. On the contrary, our discriminative model can benefit from the adaptation capability of generative models so the robustness is achieved.
- Wels et al. (2011) also propose a hybrid method for tissue segmentation. However, their approach is sequential: the discriminative model serves for

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

initializing and constraining the subsequent fitting of the generative model, which is carried out with an expectation maximization algorithm. The approach is validated on 38 scans from the same dataset (Internet Brain Segmentation Repository) using only three classes (WM/GM/CSF). It is unclear how this method would generalize to a higher number of brain structures, since the application of discriminative models at the structure level is more challenging than at tissue class level (i.e., just three classes).

- Fan et al. (2007) use a deformable template (generative) to estimate a rough region of each tissue, and a discriminative classifier (SVM) is trained on all tissue volumetric measures.
- Wang et al. (2011) propose a two-stage classifier wrapper for adapting different labeling protocols. They use the second stage classifier to learn the systematic errors from the first stage. As in Fan et al., the classifier independently refines an initial solution provided by the generative model.

Compared with these approaches, the key aspect that differentiates our algorithm is the use of an adaptive, augmented feature space that allows us to effectively fuse generative and discriminative models (as described in section “Model Fusion)”, rather than simply cascading them.

Model Description

In this section, a general formulation of the segmentation problem is provided. We first compare the generative and discriminative approach to the segmentation problem. The comparison reveals their complementary properties, and inspires our hybrid method. The idea is to first use a deformable template (the generative component, described in section “Deformable Templates Guided by Generative Models”) that produces a first estimate of the segmentation for an input volume. This rough estimate is then used (as described in in section “Model Fusion”) to: (1) calculate the regional statistics of the input volume for local (discriminative) feature normalization, and (2) derive local features such as region boundaries and label priors. These features, in combination with local appearance-based, discriminative features, give an augmented set that yields a rich description of the data by integrating information from the generative and discriminative sides. This augmented feature set will be used to train a classifier as described in section “Learning and Using Fusion Models”.

Problem Formulation

Our goal is to segment a given 3D volume/image $\mathbf{V}$ into K anatomical structures, where K is a fixed number. A training set of N volumes with their corresponding annotations (no less than K labels) is assumed to be available, and we denote this set as $S = \left\{(\mathbf{V}_1^{tr}, A_1^{tr}), \cdots (\mathbf{V}_N^{tr}, A_N^{tr})\right\}$, where $\mathbf{V}_i^{tr}$ and A_i^{tr} are the ith training volume and its corresponding annotation. We represent each structure by a region R_i. A segmentation is denoted as

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$W = \{(R_i, \Theta_i), i = 0 \ldots K\}, \quad (1)$$

where R_0 refers to the background region, and each R_i consists of all the voxels of the i^{th} anatomical structure.

These regions are disjoint and they cover the entire volume: $\cup_{i=0}^{K} R_i = \Lambda$, where Λ defines the 3D lattice of the input $\mathbf{V}$, and $R_i \cap R_j = \emptyset$, $\forall i \neq j$. Θ_i is a vector that includes the model parameters for the appearance and shape of region i.

If we define $p(\mathbf{V}(R_i)|R_i, \Theta_i)$ as the likelihood of the volume confined in region R_i under model parameters Θ_i, the optimal solution in a Bayesian framework can be obtained as:

$$W^* = \underset{W}{\operatorname{argmax}}\, p(W \mid \mathbf{V}) = \underset{W}{\operatorname{argmax}}\, p(\mathbf{V} \mid W)p(W) \quad (2)$$
$$= \underset{W}{\operatorname{argmax}} \prod_{i=0}^{K} p(\mathbf{V}(R_i) \mid R_i, \Theta_i)p(R_i)p(\Theta_i),$$

where $p(R_i)$ is the probability of the shape prior, whereas $p(\Theta_i)$ puts a prior on the parameters and is usually assumed to be flat i.e., $p(\Theta_i) \propto 1$. Moreover, we have assumed that the shapes of the different regions are independent, which allows us to write $p(W) = \Pi p(R_i)$. Whereas a more faithful model would consider dependencies between the shapes of the structures, this common assumption greatly simplifies both the training of the model and the inference in the Bayesian framework. We further assume that the intensity inhomogeneity of a region is smooth and small enough (Leemput et al. 1999).

In general, independent identical distribution (i.i.d.) assumptions are made in the likelihood function, and the appearance of each structure is approximated by a Gaussian model (Fischi et al. 2004; Pohl et al. 2006). Let $G(\cdot;\Theta_i)$ denote a Gaussian distribution parameterized by Θ_i and let v_j be the intensity value of voxel j. The likelihood then can be represented as:

$$p(\mathbf{V}(R_i) \mid R_i, \Theta_i) = \prod_{\forall j \in R_i} G(v_j; \Theta_i),$$

where Θ_i contains the mean and standard deviation of region i: $\Theta_i = \{\bar{v}_i, \sigma_i\}$. As argued in Tu et al. (2008), using only generative models with i.i.d. assumptions is often too simplistic to fully account for the realistic textures of MRI data. This insufficiency, especially between structural boundaries, will be addressed by the discriminative models in our method.

For a discriminative approach, there is no explicit data parameter estimated for each input volume $\mathbf{V}$; the model is instead learned in the form of a classifier derived from a training dataset. Thus, the solution vector by a discriminative model becomes

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$W_R = \{R_i, i = 0 \ldots K\}. \quad (3)$$

We use $|\mathbf{V}|$ to represent the total number of voxels in $\mathbf{V}$ and l_j to denote the label assigned to voxel j. R_i is therefore the set of all $l_j = i$. A discriminative classifier directly computes the class label at a voxel j which has the maximum class posterior, which is based on the local features computed on the sub-volume $\mathbf{V}(N_j)$ centered at j:

$$W_R^* \equiv (l_j^*, j = 1.. \mid \mathbf{V} \mid) = \arg\max_{W_R} \prod_{j=1}^{\mid \mathbf{V} \mid} p(l_j \mid \mathbf{V}(N_j)), \quad (4)$$

where the label of voxel j maximizing the equation is denoted as l_j^*.

If we compare the solution vectors W and W_R in Eqs. 2 (generative) and 4 (discriminative), we make two observations. First, generative models explicitly estimate the data parameters and thus are adaptive to the input. Second, discriminative models can efficiently capture complex local image statistics by combining many low- and mid-level features. As discussed above, generative models make simplistic assumptions for the likelihood term modeling the local appearance, whereas discriminative models struggle capturing the information from larger regions. Therefore, in this study we will use deformable templates guided by generative models (as described in section “Deformable Templates Guided by Generative Models”) and seek to fuse them with the discriminative model in Eq. 4 (as described in section “Model Fusion”).

Deformable Templates Guided by Generative Models

For the generative model of image intensities we adopt a Gaussian mixture model due to its modeling capability (Fischl et al. 2002; Yang et al. 2004; Pizer et al. 2003) and relatively low computational complexity. Let $\Theta = \{\Theta_i, i = 0, \ldots, K\}$, we introduce the weights of Gaussian components so the parameters for each region are given by:

$$\Theta_i = \left\{\left(\bar{v}_i^{(m)}, \sigma_i^{(m)}, \beta_i^{(m)}\right), m = 1, 2\right\},$$

where $\bar{v}_i^{(m)}, \sigma_i^{(m)}$, and $\beta_i^{(m)}$ are respectively the mean, standard deviation and weight of Gaussian component m of region R_i. In this paper, we assume two components $m = 1, 2$ for each model, which is empirically sufficient to describe the appearance of the anatomical regions. The parameters of these Gaussian components are obtained with a expectation maximization (EM) algorithm (Calinon et al. 2007). According to this model, we have the following likelihood function of voxel j in region R_i:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$p(v_j;\Theta_i)=\sum_{m=1}^{2}\beta_i^{(m)}G\left(v_j;\bar{v}_i^{(m)},\sigma_i^{(m)}\right), \quad (5)$$

Given a volume **V**, we seek to minimize an energy function under the Bayesian formulation of Eq. 2 with flat p(Θ_i):

$$E(W_R,\Theta,\mathbf{V})=\sum_{i=0}^{K}-\log\, p(\mathbf{V}(R_i)\mid R_i,\Theta_i)-\log\, p(R_i)-\log\, p(\Theta_i) \quad (6)$$

$$\approx\sum_{i=0}^{K}\sum_{j\in R_i}\left\{-\log\, p(v_j;\Theta_i)+\kappa\sum_{j'\in N_j}\delta\left(l\left(j'\right)\neq l(j)\right)\right\},$$

where the first term assumes the mixture model in Eq. 5 and the second term is a Markov Random Field prior for $p(R_i)$ that encourages smooth region boundaries/surfaces; using other priors is also possible. N_j is the set of neighboring voxels of j and $l(j')$ and $l(j)$ are respectively the region labels of j' and j; $\delta(\cdot)$ is Kronecker's delta and κ is a constant that balances the weight of the smoothness prior versus the likelihood of the intensities. We assume $p(\Theta_i)\propto 1$ so this term can be omitted. Such a flat prior implies that, a priori, we do not prefer any values of the Gaussian parameters over others. In other words, we assume no prior knowledge on the intensities of the image.

An estimate of W_R and Θ can be obtained by minimizing $E(W_R,\Theta,\mathbf{V})$:

$$\widehat{W}=\left(\widehat{W}_R,\widehat{\Theta}\right)=\underset{W_R}{\operatorname{argmin}}\,E(W_R,\Theta,\mathbf{V}). \quad (7)$$

Starting from an initial solution (a deformed template containing a volume V_a and its label annotation A_a), we minimize the energy $E(W_R,\Theta,\mathbf{V})$ in Eq. 6 using a region competition algorithm (Tu et al. 2008; Zhu and Yuille 1996). Henceforth, we denote this algorithm as *gmDT* (generative model based on a deformable template).

To avoid the initial solution is biased, we use the set of training volumes with their corresponding labels, $S=\left\{\left(\mathbf{V}_1^{tr},A_1^{tr}\right),\cdots\left(\mathbf{V}_N^{tr},A_N^{tr}\right)\right\}$, to generate the template denoted as (V_a, A_a). We use $D(\mathbf{V},\mathbf{V}_n)$ to denote the dissimilarity between **V** and $\mathbf{V}_n^{tr}$ after applying a linear transformation (in our case, computed with AIR (Woods et al. 1993)). The learned template volume $\mathbf{V}_a$ minimizes the total dissimilarity with all other volumes in the training set:

$$\mathbf{V}_a=\arg\min_{\mathbf{V}\in S}\sum_{n=1}^{N}D\left(\mathbf{V},\mathbf{V}_n^{tr}\right), \quad (8)$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The corresponding manual annotation of V_a, A_a, is used as the initial W_R in Eq. 7. Then, the obtained segmentation estimated by *gmDT*, $\hat{W}_R$, is an approximation to the optimal solution. A byproduct of *gmDT* is $\hat{\Theta}$, the parameter estimates.

Model Fusion

Next, we discuss the discriminative model in our method, which is used to incorporate the adapted information from *gmDT*. For a purely discriminative model, we denote the total number of candidate features by B and the k^{th} feature for voxel j computed on volume $\mathbf{V}(N_j)$ by $f_{d,k}(j)$. The discriminative feature vector $\mathbf{F_d}(j)$ of each voxel j can be written as:

$$\mathbf{F_d}(j) \equiv \left[f_{d,1}(\mathbf{V}(N_j)), \cdots, f_{d,B}(\mathbf{V}(N_j)) \right] \quad (9)$$

A discriminative classifier, e.g., boosting (Freund and Schapire 1997), selects a number of informative features (typically a couple of hundred) from $\mathbf{F_d}(\cdot)$ and fuses them with appropriate weights. The training process is driven by the minimization of the classification error in the labeled training data and the generalization power of the classifier (VC dimension, Vapnik 1982). Hence, the quality of a trained discriminative classifier is greatly determined by the effectiveness of its feature set.

To achieve enhanced robustness, the basic idea here is to integrate the adaptiveness and the fusion capability respectively from generative and discriminative models. This is done by augmenting the feature vector $\mathbf{F_d}$ with ($\hat{W}_R$, $\hat{\Theta}$) from the deformable template. This way, we achieve robustness against intensity variations while we equip $\mathbf{F_d}$ with structure-adapted features.

Using $\hat{\mathbf{W}}_\mathbf{R}$ for Intensity Normalization—Features computed directly from $\mathbf{V}(N_j)$ are often sensitive to geometrical and intensity variations, but $\hat{\Theta}$ from Eq. 7 can then be used to normalize $\mathbf{V}$. We denote the normalized volume as $\mathbf{V}_{\hat{\Theta}}$. The new, augmented feature vector $\mathbf{F}(j)$ is:

$$\mathbf{F}(j) \equiv \left[F_{\hat{W}_R}(j), f_{d,1}(\mathbf{V}_{\hat{\Theta}}(N_j)), \cdots, f_{d,B}(\mathbf{V}_{\hat{\Theta}}(N_j)) \right], \quad (10)$$

Comparing Eq. 10 with Eq. 9, $f_{d,k}(\mathbf{V}_{\hat{\Theta}}(N_j))$is computed on the normalized volume $\mathbf{V}_{\hat{\Theta}}$, instead of $f_{d,k}(\mathbf{V}(N_j))$, for a voxel j. $F_{\hat{W}_R}(j)$ represents the augmented features based on $\hat{W}_R$ that will be discussed below.

Normalization is achieved by intensity correction based on matching the intensity of the regions to those from the template volume (Hou 2006). Specifically, we search for the linear transform that best matches the intensities in a least squares sense, a problem that can be solved with standard techniques (Tibshirani 1996).

Augmenting Atlas Features from $\hat{\mathbf{W}}_\mathbf{R}$—From $\hat{W}_R$ (given by *gmDT*), we have an estimated region label for each voxel j. To differentiate this estimated label from l_j in Eq. 4,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

we use $\hat{l}_j$ to denote it. From $\hat{W}_R$, the displacement vector of a voxel j to the centroid of structure k, $d_{\hat{R}_k}(j)$, can be calculated. This displacement vector is a spatial feature which is more adaptive than the absolute coordinates. Similarly, we can compute the signed distance function of each voxel j with respect to the estimated region boundary of each anatomical structure. The signed distance $s_{\hat{R}_i}(j)$ of voxel j to the boundary of region $\hat{R}_i$ is:

$$s_{\hat{R}_i}(j) = \begin{cases} + \min_{j' \in \hat{C}_i} d(j,\ j')\ \ if\ \ \hat{l}(j) = i \\ - \min_{j' \in \hat{C}_i} d(j,\ j')\ \ otherwise \end{cases}, \quad (11)$$

where $d(j, j')$ is the distance between j to any point j' on the region boundary $\hat{C}_i$. The positive/negative sign indicates that j is inside/outside $\hat{R}_i$. Now our augmented feature vector becomes:

$$F_{\hat{W}_R}(j) = \left(d_{\hat{R}_0}(j), \cdots, d_{\hat{R}_K}(j), s_{\hat{R}_0}(j), \cdots, s_{\hat{R}_K}(j) \right), \quad (12)$$

Henceforth, we denote these features derived from $\hat{W}_R$ as “atlas features”. They correspond to $F_{\hat{W}_R}(j)$ in Eq. 10.

Learning and Using Fusion Models

Here we aggregate all the components in section “Model Description” to define the training and classification (testing) stages of our method. An atlas (template) will first be selected among the training data as the template. As described in section “Deformable Templates Guided by Generative Models”), the template will be guided by *gmDT* to adapt to the input volumes. Though we know the true labels of the training data, *gmDT* is applied (normalization and atlas features) in training the discriminative models so the resultant classifiers can model the estimates from *gmDT* in the test stage.

Atlas Selection and Feature onstitution

In the training stage, we have the set of training volumes with their corresponding labels $S = \left\{ \left(\mathbf{V}_n^{tr}, A_n^{tr} \right), n = 1..N \right\}$, which have been normalized to the same scale and properly preprocessed. A template, $(\mathbf{V}_a, A_a)$, can be learned based on Eq. 8.

For our model, we will need to learn the uncertainty from *gmDT* before training the discriminative models. Using A_a as the initial labeling, we perform *gmDT* on the rest of training volumes in S. For each training volume $\mathbf{V}_n^{tr}$, an estimated segmentation $\hat{W}_n^{tr}$ is obtained from *gmDT*. On the other hand, we define the feature set for the discriminative classifiers, $\mathbf{F_d}(\cdot)$, as local features such as gradients, curvatures, and Haar-like responses at various spatial scales (approximately 5,000 in total in this paper, but our method is not restricted to the specific $\mathbf{F_d}(\cdot)$). Since the training volumes are in the same size, these features can be computed directly on a pre-defined sub-window (of size 11 × 11 × 11)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

centered at the target voxel. A detailed discussion of how these features are computed can be found in Tu et al. (2008).

We use the estimated $\hat{W}_n^{tr}$ to: (1) normalize the intensity of $\mathbf{V}_n^{tr}$ to $\mathbf{V}_a$ as in section "Model Fusion" and then compute $[f_{d,1}(\mathbf{V}_{\hat{\Theta}}(N_j)), \cdots, f_{d,B\approx 5000}(\mathbf{V}_{\hat{\Theta}}(N_j))]$ on the normalized intensity volume; and (2) compute the atlas features $F_{\hat{W}_R}(j)$ as in Eq. 12. Combining the two sets yields the augmented feature set (about 6,000 features), $\mathbf{F}(j)$, for each voxel j in the training volume $\mathbf{V}_n^{tr}$.

In short, *gmDT* is also applied to the training volumes. The role of the ground truth labels in the training stage is to train the discriminative, supervised classifiers; they do not participate before computing the augmented features.

Integration within a Discriminative Framework

Once all the features $\mathbf{F}(j)$ are computed, we train a classifier upon the training set

$$\aleph = \{(l(j), \mathbf{F}(j)), j = 1..T\},$$

where $l(j)$ is the true label for voxel j, and T is the total number of voxels in all the training volumes.

A learning algorithm either directly combines all the features in $\mathbf{F}(j)$ like SVM (Vapnik 1998), or selects a set of features out of $\mathbf{F}(j)$ such as boosting (Freund and Schapire 1997) and random forests (Breiman 2001). Either way, features are combined into the classifier in order to minimize the training error. When training our model, no preference was given to features derived from the atlas over those computed from image intensities (e.g., Haar-like), and vice versa. Here we adopt the auto-context algorithm using a cascade of PBT (Probabilistic Boosting Tree) classifiers (Tu and Bai 2010) as the discriminative model, which performs feature selection and fusion by exploring a high-dimensional feature space. Note that, in addition to the features in $\mathbf{F}(j)$, auto-context (Tu and Bai 2010) itself is an iterative method that incorporates contextual information into the classification by augmenting the feature space as follows. In a first iteration, the classifier is trained on the available set of features. In subsequent iterations, the label posteriors (as estimated by the current classifier) at a fixed set of shifted locations are added to the feature space, implicitly capturing the shape of the structures to segment. A summary of the algorithm is described in Fig. 1; the reader is referred to the original paper (Tu and Bai 2010) for further details.

Nevertheless, the key of our proposed method is the augmentation/normalization of features, which implicitly fuses the generative and discriminate aspects of the model. It is not tied to any specific choice of classifier so one could also use boosting, random forests, or any probabilistic classifiers as the discriminative classifier.

Once a classifier has been trained on the training set $\aleph = \{(l(j), \mathbf{F}(j)), j = 1..T\}$, we can use it to estimate $p(l|\mathbf{F}(i))$ for a given test voxel i. A test volume is required to have the same

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

preprocessing steps and resized such that all features can be correctly computed. Each test voxel will be assigned to the label that maximizes the probability: $l_i^* = \arg\max_l p(l \mid \mathbf{F}(i))$.

We summarize the training algorithm in Fig. 2 and illustrate it in Fig. 3. In Fig. 3, the identical template, ($\mathbf{V}_a$, A_a), will be used in both training and testing. The testing procedure consists of three stages analogous to those in the training procedure, except the last stage: performing classification using the trained classifier.

Experiments

To examine the effectiveness of the proposed algorithm, we perform a thorough empirical study using four MRI T1-weighted brain datasets and compare our method against the state-of-the-art systems. We focus on sub-cortical structures due to their popularity in the literature. Besides comparing our method with the two components in our method, *gmDT* and *DM*, we also include the methods developed by other researchers to show the integration can achieve better performance in most scenarios.

We will demonstrate the performance of the integrated method by four parts: (1) Increased importance of the adapted atlas features using the fusion mechanism; (2) Comparison using the same dataset for training and testing (intra-dataset). This scenario is very suit for *DM* and the proposed method is comparable to it; (3) Comparison using different datasets for training and testing (inter-dataset). This scenario is favored by *gmDT* and our method achieved better results; (4) Performance on longitudinal data, which shows the potential of our method to capture the morphological changes by the same subject. Though the fact that *DM* tends to fail in the last two parts is known, we still show its quantitative results for completeness.

Experimental Setup

In this section, we elaborate the datasets used in our four parts of experiments, the three main algorithms to compare, the pre-processing steps we applied to the heterogeneous datasets, and the measures we used for our comparisons.

T1 MRI Datasets—All the datasets used in our experiments are the following (the suffix indicates the number of volumes in the dataset):

(1) *IBSR18*: We use the 18 scans with 84 manually annotated structures from the Internet Brain Segmentation Repository (IBSR)[1] (we do not include the 120 cortical/subcortical parcellations as 84 structures are sufficient in our experiment). All volumes were scanned at 1.5T. This dataset has been extensively utilized as a benchmark for evaluation in various papers.[2]

(2) *LPBA40*: The LONI Probabilistic Brain Atlas dataset (LPBA) (Shattuck et al. 2008) contains volumes from 40 healthy subjects with 56 anatomical structures manually annotated. These structures include both cortical and subcortical regions. All subjects were scanned with a GE Signa 1.5T systems with a SPGR sequence.

(3) *LONI28*: 28 scans from normal subjects were acquired on a GE Sigma 1.5T scanner with a SPGR sequence. Eight subcortical structures (left and right hippocampus, putamen, caudate nucleus, and lateral ventricle) were manually delineated by neuroanatomists.

(4) *ADNI980*: The Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jack et al. 2008) was launched in 2003 by government, private pharmaceutical companies and non-profit organizations. The goal of ADNI has been to develop measures for the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). ADNI is the result of efforts of many co-investigators from a broad

[1] http://www.cma.mgh.harvard.edu/ibsr
[2] http://www.cma.mgh.harvard.edu/ibsr/PubsUsingIBSR.html

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

range of academic institutions and private corporations, and the data keeps growing by its follow-up projects. Readers may see www.adni-info.org for more information. In this study, 490 pairs of brain scans (980 volumes in total) were selected from ADNI. They correspond to 490 subjects who are known to be one of the three groups: elderly controls, mild cognitive impairment (MCI), or Alzheimer's Disease (AD). For each subject, two volumes were acquired 12 months apart for longitudinal analysis. This dataset is particularly challenging because the subjects were scanned at different sites with different scanners. Therefore, these volumes display high variability stemming from intrinsic longitudinal changes of subjects and scanning configurations. More details of the various scanning protocols can be found in http://adni.loni.ucla.edu/research/protocols/mri-protocols/.

Note that not all the datasets have manual delineations of all subcortical structures. We summarize the main characteristics of these four datasets, as well as the role they play in our experiments, in Table 1.

Algorithms to Compare—We adopt region competition (Tu et al. 2008; Zhu and Yuille 1996) as the *gmDT* process which iteratively minimizes the energy $E(W_R, \Theta, \mathbf{V})$ in Eq. 6. We choose region competition due to its simplicity and effectiveness. The template (V_a, A_a) is learned according to Eq. 8, using Mattes mutual information as the dissimilarity function. The region competition process takes about 5 ~ 15 minutes to reach a steady state (no more change of labels, or fluctuation of labels) of surface evolution. It runs typically for less than 15 iterations. Other approaches such as the level set methods (Chan and Vese 2001; Yang et al. 2004) could also have been used to perform energy minimization in a similar manner.

On the other hand, we use the auto-context algorithm (Tu and Bai 2010) with PBT as the baseline discriminative classifier. Henceforth, this discriminative model is denoted as *DM*. Its running time is 10 ~ 20 minutes, depending on the number of structures for segmentation.

Comparisons between our method, *gmDT* and *DM* demonstrates the advantages due to the integration. We also list the measures from literatures for the same T1 MRI dataset and structures if available.

Pre-Processing—Before applying the three algorithms for comparison, these heterogeneous datasets several preprocessing steps. We eliminate the dominant spatial disparity between an input volume and an atlas by the following sequence of preprocessing steps: (1) re-orientation using AIR 2.5 (Woods et al. 1993) (if the two volumes were at different orientations); (2) skull stripping using BET in FSL 4.0.3 (Smith 2002); (3) a 12-parameter global affine registration using AIR 2.5; and (4) a diffeomorphic registration, SyN, from ANTS 1.9 (Avants et al. 2008).

We chose SyN as our non-linear method stage due to its speed and high accuracy (Klein and et al. 2009). We use the following settings: three resolution levels (30x50x5 iterations), step-length 0.15, probability mapping (PR) with 4 mm radius as cost function, and regularization with a Gaussian filter with standard deviation 3 mm. These parameters are obtained empirically and they provide sufficient spatial alignments for *gmDT* across the four test datasets. Under these settings, the SyN registration between an atlas and an image can be done under 30 minutes. Steps prior to SyN can be done in 5 minutes. Once the segmentation result is obtained, each preprocessing step is reverted to map the result back to the original space.

Measures for Comparison—The main evaluation measure used here is the Dice overlap, $Dice = \frac{2 * |L \cap S|}{(|L| + |S|)}$, where L and S are the sets of voxels manually annotated and those

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

automatically segmented. Precision and recall rates are also used in Table 6, where $Precision = |L \cap S|/|S|$ and $Recall = |L \cap S|/|L|$. Another popular measure in literature, Jaccard coefficient, can be directly calculated from Dice: $Jaccard^{-1} = 2Dice^{-1} - 1$. In the inter-dataset tests, as our method and *gmDT* are adaptive, we further compare them in surface consistency. Hausdorff distances (Loncaric 1998) and the Mean distances (Yang et al. 2004) between two sets of surface voxels are measured:

$$H(A,B) = \max_{a \in A} \min_{b \in B} D(a,b),\ M(A,B) \quad (13)$$

$$= \sum_{a \in C_A} \min_{b \in C_B} \frac{D(a,b)}{|C_A|},$$

where A and B are sets of voxels. $H(\cdot)$ and $M(\cdot)$ are respectively the Hausdorff distance and the Mean distance. D is the underlying distance metric, which is usually the Euclidean distance or the Manhattan distance. C_A is the surface of segment A. Note that both the directed Hausdorff distance and the Mean distances are not symmetric. We will use H(A,B), H(B,A), and one direction (segmented to ground truth) of the Mean distances in our evaluation.

Importance of Atlas Features

To demonstrate the importance of using atlas features based on deformable templates, we compare the features used in two models trained to segment the 56 brain structures of LPBA, one with a fixed annotation A_a (no adaptation to the input volume) and another with an adapted atlas based on *gmDT*. The feature pool is dominated by Haar-like responses due to their effectiveness to describe the appearance at different spatial scales. Derivative features perform similarly to Haar, but are limited to eighteen local derivatives in the x, y, and z directions. Features in the column “Others” include intensities, gradients, and curvatures. As mentioned in section “Integration within a Discriminative Framework”, we perform feature selection when training the Probability Boosting Tree (PBT). For each node in PBT, a limited number of features are chosen to form a decision criterion (a boosting classifier) such that the classification error is minimized for the training data arriving at this node. Giving the same complexity of the trained classifiers (the same tree depth and the same number of features used at each tree node), we observe in the table that more atlas features are selected when using an adapted atlas rather than a fixed one; the percentage of the selected atlas features increases from <3 % to 6 % (Table 2).

Intra-Dataset Evaluation

In this section, we evaluate our algorithm using training and test data from the same dataset, which is common in brain image segmentation. IBSR18 and LPBA40 are included in this experiment.

(1) *IBSR18*: A total of 14 subcortical structures including the left and right lateral-ventricles are evaluated. Due to the relatively few number of volumes in IBSR18, we perform a threefold cross-validation, i.e., six volumes in each fold are used as the test set, and the other twelve volumes are used to train our model.
We use the box plots in Fig. 4 to compare the three methods and Table 3a to lists the Dice overlaps produced by DM, pure *SyN*, *gmDT*, and our method. The measures of *SyN* were included to clarify the improvements between it and our generative process.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

In Table 3a, the row *SyN* lists six of the intermediate DICE overlaps before using *gmDT*. *SyN* performs a diffeomorphic transform mainly adapted to larger spatial variations of the whole brain so it can reduce location errors. Based on its rough results, *gmDT* utilized the relatively reliable statistics of each structure so the DICE values can be boosted at least 0.039(Thalami). Therefore, the use of *SyN* in our system is to reduce the risk of failure in *gmDT*; it is still a preprocessing step instead of a main factor in the whole our method. The data in IBSR18 display larger variations in appearance and shape, and therefore, DM trained upon just 12 images in each fold could not generate satisfactory results. In this scenario, *gmDT* showcases its superior adaptiveness. From Fig. 4, we observe that the advantage of *gmDT* is inherited by our fusion method but the similar and better measures.

A visual example of the extracted structures by DM, *gmDT* and our method is shown in Fig. 5. Although not all of the 18 structures can be shown in a single 2D slice, better matched caudate nuclei, thalami, and putamens by our algorithm due to the fused information can still be observed. Table 3b lists several published reports on IBSR18, and our measures are at the top on the seven types of gray matter structures, with particularly significant improvements for the pallidum and amygdala. The results are close to Khan et al. (2009) which is based on a computationally expensive multi-atlas approach. Their method outperforms ours for the ventricles (0.85 versus 0.80), but is inferior for the pallidum (0.72 versus 0.81) and amygdala (0.66 versus 0.73).

(2) *LPBA40*: In the LPBA40 dataset Shattuck et al. 2008, 40 subjects with fifty-six anatomical structures, both cortical and sub-cortical, were manually delineated for each T1-weighted MRI volumes. We randomly choose 25 volumes for training and use the remaining 15 for testing.

To compare our algorithm against *gmDT* and *DM*, we use Fig. 6 and the detailed measures in Table 4 for a comprehensive comparison. From Fig. 6, we observe that fusing the two models in our method achieves higher average Dice coefficients than the two baseline methods. Our method gives better measures on L/R hippocampus (p-value=0.017/0.003) than both the discriminative and the generative models, but the measures of caudate nuclei are worse than DM due to the considerably degraded performance of the generative model. Reasons for this could be (1) the intra-dataset test of LPBA40 containing smaller data variance is favored by DM, and (2) the boundaries between the caudate nuclei and the ventricles in LPBA40 are obscured so *gmDT* could be misled. The weaknesses of *gmDT* in this scenario are the inferior DICE scores and quartile positions. These are all improved by our fusion method due to the corrections in the discriminative framework, especially in the left/right caudates and the left putamen. The difference in the average Dice overlap between our method and *gmDT* is relatively small (~ 1 %), but still very statistically significant by the p-values shown in Table 4.

In summary, the two intra-dataset tests show the proposed fusion method successfully combine the advantages of *DM* and *gmDT*, and its performance is at least comparable to *DM*.

Robustness Across Different Datasets (Inter-Dataset)

In the previous experiments, we trained and tested our method on volumes from the same dataset. To build a practical system dealing with clinical data, it is important to test its robustness on a large number of volumes from various sources.

We use the same model trained from IBSR18 as in section “Intra-Dataset Evaluation”(1). The test datasets include LPBA40 and LONI28. Both of them have manual annotations of subcortical structures (caudate nuclei, putamens, and hippocampi for both sets, lateral ventricles only for LONI28). These structures are all covered by the annotations of IBSR (see section “Intra-Dataset Evaluation”(1). We choose these subcortical structures because: (1) they are frequently used to evaluate automatic segmentation methods; and (2) they are very relevant in neuro-image studies of diseases such as Alzheimer’s and Parkinson’s. To keep the results comparable with other literatures on the same test datasets, we chose not to re-annotate the structures but keep the annotations as they were. In the following experiments, *DM*, *gmDT*, and our algorithm are trained from the same training data. The other methods (FreeSurfer and FSL) are off-the-shelf systems without any parameter

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

adjustments. Although the difference among the three protocols introduce intrinsic Dice errors, they still share a high proportion of common structure regions and the relative performance between the three models trained by the same data can be observed.

In addition to the quantitative analyses on the two test datasets, we also test the model trained on LPBA40 - section "Intra-Dataset Evaluation"(2) - and show visual examples on ADNI980, LPBA40 and LONI28 at the end of this section.

(1) *LPBA40*: Table 5 gives the Dice measures of three types of subcortical structures in LPBA40. All values are the averages of the corresponding measures of the left and right structures. Our method demonstrates the best performance among all five competing algorithms. Compared with the intra-dataset results in Table 4, noticeable degradation of the Dice overlaps is found in hippocampus among all methods in Table 5. This is because the subiculum region of hippocampus in LPBA40 was annotated differently from IBSR and many other datasets.
The improvement in surface distance measures given by our method is shown in Figs. 7 and 8. We denote the automated segmented result as W^* and the ground truth as W. Our method achieves consistent improvement over *gmDT*. According to the p-values, the mean distance measures show especially significant improvement on the left caudate (p-value=0.0010), left putamen (p-value=0.0041), left hippocampus (p-value=0.0003), and right putamen (p-value=0.0326). The box plots in Fig. 9 confirm the overall better performance of our method in terms of Dice overlap.

(2) *LONI28*: The segmentation accuracy in eight subcortical structures on LONI28 (Left/Right Hippocampi, Left/Right Caudate Nuclei, Left/Right Putamens, and Left/Right lateral ventricles) by the model trained from IBSR18 is shown in Table 6. The results from Freesurfer and Hybrid are from Tu et al. (2008). The Hybrid model is trained by 14 LONI28 volumes and the measures are calculated from the other 14 test volumes. Although our model is trained by IBSR18, our method provides the best F-values in five structures as well as the averages in the last column. This shows the robustness of our method when segmenting the subcortical, gray matter structures in a different dataset.
The Dice overlaps in Table 6 show similar performances by *gmDT* and our method. Surface distance measures are showed in Figs. 10 and 11. Our method gives smaller Mean and Hausdorff distances, while *gmDT* provides smaller $H(W, W^*)$. We notice that $H(W^*, W)$ yields a significant improvement on most structures (LH, RH, LC, RC, LP, RP). From the four tables, we observe that W^* produced by our method has more consistent surface with W than *gmDT* (better $H(W^*, W)$ and smaller Mean distance).
The improvement produced by our method can also be observed in the box plots in Fig. 12. Our method not only increases the average Dice coefficient in all the tested structures (higher average values), but the worst cases have all improved.

(3) *56 structures on ADNI980, LONI28, and IBSR18*: In addition to the two subcortical tests on our IBSR18 subcortical model, we used the fifty-six structure model trained upon LPBA40 in section "Intra-Dataset Evaluation"(2) to perform cortical and subcortical segmentation on three datasets: ADNI980, LONI28, and IBSR18. Figure 13 shows a number of 2D MRI slices together with their segmentation results. We see from the figure that the intensity patterns and textures are quite different for these datasets. Even within the same dataset, ADNI980, the MRI slices show large variation since not all of them were acquired with the same scanner. However, despite such a high degree of variability in the data, the segmentation results are mostly satisfactory.

By the first two parts of tests containing disparate training and test datasets, we see the proposed fusion method can achieve the highest compatibility between different protocols, all by experts, than *DM* and *gmDT* given the same training set. The third test further shows its robustness to large data variation. Both properties are important when applying the proposed method as a common brain segmentation tool.

Performance in Longitudinal Studies

Section "Robustness Across Different Datasets (Inter-Dataset)" demonstrated the effectiveness of the proposed algorithm for segmenting structures from several datasets with high variability in anatomy and image intensities. In this section, we show the results of our algorithm on data in a longitudinal study where the main source of variation is the temporal changes within the same subject. Since the results from Hua et al. 2009 indicate the significant atrophy of hippocampi in both MCI and AD cases, and the scans from ADNI980 provide the longitudinal (12 months) data of 490 subjects from AD, MCI, and normal control groups, we use this dataset as our testbed. ADNI980 volumes also display large variations between subjects due to different scanning settings; nevertheless, volumes from the same subject still share the same acquisition configurations. Although our method may need more specific design for high precision hippocampus segmentation, this experiment still shows our advantage with respect to the baseline methods in the longitudinal study.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The tested segmentation methods include DT, *gmDT*, and our algorithm. The training data contains volumes from 7 AD, 7 MCI, and 7 control subjects in ADNI with manual annotations of hippocampi by an expert (these subjects are not part of the 490). The measures listed in Table 7 include the average volume of baseline, the average volume after 12 months, the volume difference between the two times, and the percentage of volume loss.

All the measures obtained by a discriminative approach (Tu and Bai 2010) are listed in Table 7a and those by *gmDT* are listed in Table 7b. As DM is not adaptive, the coverage of the segmented region by DM is conservative if the test hippocampus has a drifted position, a shape distortion like that typically produced by AD, or a different statistical distribution of appearance. DM also fails to identify the hippocampal regions of 13 subjects, so we needed exclude their volumes from Table 7a. For the rest of the scans, the averages of volume change and percentage of loss show noticeable differences between groups. However, their standard deviations are relatively large, indicating that the measures from the direct discriminative model are not stable for the three groups. In Table 7b, the longitudinal differences between AD and normal are not fully demonstrated because *gmDT* (with a simple appearance model) could include more non-hippocampal regions than DM and our method. Instead, our approach gives apparent differences in all columns among the three groups; see Table 7c. The smaller standard deviations provide a better separation of the three groups compared to DM. In addition, the balanced performance among the left hippocampus, the right hippocampus and the average is evident.

We further compare the results by our method with the estimated hippocampal volume changing rates reported by Schuff et al. (2012). Their estimated rates are modeled as nonlinear curves based on manual segmentations of ADNI data. As the average ages of our three groups are 76.82(standard deviation(SD) = 6.63) for Normal, 76.05(SD = 6.67) for MCI, and 76.82(SD = 6.44) for AD, our data are mostly located in the range [70–80], where their estimated curves are still close to linear. Using the average hippocampal volume of 75 years old normal subjects as the standard level, their chart shows the estimated average volume losses as 400 mm^3 (sMCI, the subjects do not convert to AD), 650 (cMCI, the subjects would progress to AD), and 750 (AD). Our average volume of the whole MCI group is 573.40 mm_3 less than the Normal group and consistent with this estimation. However, the number of our AD group is 920.34 mm_3 which is 170 mm^3 larger than their estimation. This is a hint of what direction to work in if we want to specifically adjust our work in order to obtain precise hippocampal segmentations. Table 7 shows that our method achieves higher robustness when identifying inter- and intra- individual differences, and its potential to help indicate different pathological stages.

Conclusion and Future Work

In this paper, we have proposed a system for brain MRI image segmentation by fusing together deformable templates (generative) and informative features (discriminative). It takes the advantages of the generative model for being adaptive and the discriminative classifier for achieving classification power on high dimensional data. This approach uses a new way

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

of combining generative and discriminative models and the complementary properties between them can be efficiently exerted; the information extracted from generative models is considered as an additional channel of features for training the discriminative models. The trained models are improved in two ways: (1) The original features can be normalized according to intrinsic structural statistics. Typical methods to accommodate large variation, such as performing histogram matching (Sled et al. 1998) or extracting features invariant to intensity change (Unay et al. 2008), are only based on the statistics of whole volume data. (2) The feature set is augmented with features derived from the estimation by *gmDT*.

A thorough experimental study on T1-weighted datasets demonstrates the robustness of our algorithm. Although discriminative models can perform well if the training and the test data share the same condition of variances, better performance can still be observed by our method as our discriminative models augment the informative feature set with generative features. This advantage leads to improvement over several state-of-the-art algorithms. The inter-dataset and the longitudinal tests show the deficiency of discriminative models in practice and the necessity of introducing the generative information. Our method demonstrates both adaptiveness and precision in these challenging tests and outperforms the two direct models in region overlaps (Dice) and surface fitness. This is different from Wang et al. (2011) that the first stage classifier is considered as a black box approach. These advantages also lead to the improvement over several state-of-the-art algorithms on standard datasets such as IBSR18, LPBA40, and ADNI.

An important aspect of the proposed method is its running time, which is approximately one hour. Whether the system is practical depends on the application. In neuroimaging studies, in which research labs often spend months collecting the data, slow running times are not a problem. For instance, our method is much faster than the widely used FreeSurfer, which requires on average 12 hours to segment a single brain scan. In clinical practice, one hour might no be sufficiently fast. The bottleneck of the algorithm is, as for many other brain MRI segmentation methods, the nonlinear registration. However, the registration can be dramatically sped-up through parallelization.

An aspect of the framework that was not evaluated was its performance on multispectral data. In this scenario, the different data channels represent images acquired with different MRI contrast (Prastawa et al. 2003; Menze et al. 2010; Yang et al. 2010; Geremia et al. 2011) or even different modalities (Fitzpatrick et al. 1999; Chen and Varshney 2003). Segmentations on multispectral data have the potential to be more accurate thanks to the larger amount of information present in the different channels. In our framework, generalization to multispectral scenarios is immediate: the intensities of the additional channels are just extra dimensions of the feature vectors. Another possible line of research would be to analyze the performance of the framework using other generative or discriminative models. For instance, it would be interesting to assess whether introducing explicit shape, regional, or context information in the generative prior has a positive impact of the segmentation. Exploring all these directions remains as future work.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Information Sharing Statement

An implementation of the method is publicly available for download at the LONI[3] and NITRC [4] websites. We provide a Windows®, a Linux®, and a LONI pipeline version. The software can be used freely in research provided this paper is cited in any material using the results of their application. For other usage, contact the authors.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was funded by NSF CAREER award IIS-0844566, ONR award no. N000140910099, and NSF award IIS-1216528.

Data collection and sharing for this project was also funded by the 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, Alzheimer's Association, Alzheimer's Drug Discovery Foundation, Amorfix Life Sciences Ltd., AstraZeneca, Bayer HealthCare, BioClinica, Inc., Biogen Idec Inc., Bristol-Myers Squibb Co., Eisai Inc., Elan Pharmaceuticals Inc., Eli Lilly and Company, F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc., GE Healthcare, Innogenetics, N.V., IXICO Ltd., Janssen Alzheimer Immunotherapy Research & Development, LLC, Johnson & Johnson Pharmaceutical Research & Development LLC, Medpace, Inc., Merck & Co., Inc., Meso Scale Diagnostics, LLC, Novartis Pharmaceuticals Corp., Pfizer Inc., Servier, Synarc Inc., and Takeda Pharmaceutical Co.. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129 and K01 AG030514.

References

Akselrod-Ballin, A., Galun, M., Gomori, J., Basri, R., Brandt, A. Atlas guided identification of brain structures by combining 3D segmentation and SVM classification. Proceedings of international conference on medical image computing and computer assisted intervention (MICCAI'06), Part I; 2006. p. 209-216.

Akselrod-Ballin, A., Galun, M., Gomori, J., Brandt, A., Basri, R. Prior knowledge driven multiscale segmentation of brain MRI. Proceedings of international conference on medical image computing and computer assisted intervention (MIC-CAI'07), Part II; 2007. p. 118-126.

Aljabar, P., Heckemann, R., Hammers, A., Hajnal, J., Rueckert, D. Classifier selection strategies for label fusion using large atlas databases. Proceedings of international conference on medical image computing and computer assisted intervention (MICCAI'07), Part I; 2007. p. 523-531.

Ashburner J, Friston K. Unified segmentation. NeuroImage. 2005; 26(3):839–851. [PubMed: 15955494]

Avants B, Epstein C, Grossman M, Gee J. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis. 2008; 12:26–41. [PubMed: 17659998]

Bauer, S., Nolte, LP., Reyes, M. Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization. Proceedings of international conference on medical image computing and computer assisted intervention (MICCAI'11), Part III; 2011. p. 354-361.

[4] http://www.nitrc.org/projects/brainparser

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Bazin PL, Pham D. Topology-preserving tissue classification of magnetic resonance brain images. IEEE Transactions on Medical Imaging. 2007; 26(4):487–496. [PubMed: 17427736]

Bazin PL, Pham D. Homeomorphic brain image segmentation with topological and statistical atlases. Medical Image Analysis. 2009; 12:616–625.

Breiman L. Random forests. Machine Learning. 2001; 45(1):5–32.

Caldairou B, Passat N, Habas PA, Studholme C, Rousseau F. A non-local fuzzy segmentation method: application to brain MRI. Computer Analysis of Images and Patterns. 2011; 44(9):1916–1927.

Calinon S, Guenter F, Billard A. On learning, representing and generalizing a task in a humanoid robot. Special issue on robot learning by observation, demonstration and imitation. IEEE Transactions on Systems, Man and Cybernetics, Part B. 2007; 37(2):286–298.

Chan T, Vese L. Active contours without edges. IEEE Transactions on Image Processing. 2001; 10(2): 266–277. [PubMed: 18249617]

Chen HM, Varshney P. Mutual information-based CT-MR brain image registration using generalized partial volume joint histogram estimation. IEEE Transactions on Medical Imaging. 2003; 22(9): 1111–1119. [PubMed: 12956266]

Corso J, Sharon E, Dube S, El-Saden S, Sinha U, Yuille A. Efficient multilevel brain tumor segmentation with integrated Bayesian model classification. IEEE Transactions on Medical Imaging. 2008; 27:629–640. [PubMed: 18450536]

Du J, Younes L, Qiu A. Whole brain diffeomorphic metric mapping via integration of sulcal and gyral curves, cortical surfaces, and images. NeuroImage. 2011; 56(1):162–173. [PubMed: 21281722]

Essen DCV, Dickson J, Harwell J, Hanlon D, Anderson CH, Drury HA. An integrated software system for surface-based analyses of cerebral cortex. Journal of American Medical Informatics Association. 2001; 8(5):443–459.

Fan Y, Shen D, Gur RC, Gur RE, Davatzikos C. COMPARE: classification of morphological patterns using adaptive regional elements. IEEE Transactions on Medical Imaging. 2007; 1:93–105.

Felzenszwalb PF. Representation and detection of deformable shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2005; 27(2):208–220. [PubMed: 15688558]

Fischl B, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron. 2002; 33:341–355. [PubMed: 11832223]

Fischi B, Salat DH, van der Kouwe AJ, Makris N, Segonne F, Quinn BT, Dale AM. Sequence-independent segmentation of magnetic resonance images. NeuroImage. 2004; 23:S69–S84. [PubMed: 15501102]

Fitzpatrick J, Wang MY, Dawant BM, Maurer CRJ, Kessler RM, Maciunas RJ. Retrospective intermodality registration techniques for images of the head: surface-based versus volume-based. IEEE Transactions on Medical Imaging. 1999; 18(2):144–150. [PubMed: 10232671]

Frackowiak, R., Friston, K., Frith, C., Dolan, R., Price, C., Zeki, S., Ashburner, J., Penny, W. Human brain function. New York: Academic Press; 2003.

Freund Y, Schapire R. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Comparative and System Sciences. 1997; 55(1):119–139.

Geremia E, Clatz O, Menze BH, Konukoglu E, Criminisi A, Ayache N. Spatial decision forests for ms lesion segmentation in multi-channel magnetic resonance images. NeuroImage. 2011; 57(2):378–390. [PubMed: 21497655]

Geremia, E., Menze, BH., Clatz, O., Konukoglu, E., Criminisi, A., Ayache, N. Spatial decision forests for MS lesion segmentation in multi-channel MR images. Proceedings of international conference on medical image computing and computer assisted intervention (MICCAI'10), Part I; 2010. p. 111-118.

Gouttard S, et al. Subcortical structure segmentation using probabilistic atlas priors. Medical Imaging. 2007; 6512:1–11.

Heckemann R, Hajnal J, Aljabar P, Rueckert D, Hammers A. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage. 2006; 33(1):115–126. [PubMed: 16860573]

Holub A, Welling M, Perona P. Hybrid generative-discriminative visual categorization. International Journal of Computer Vision. 2008; 77(1–3):239–258.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Hou Z. A review on MR image intensity inhomogeneity correction. International Journal of Biomedical Imaging. 2006; 2006:1–11.

Hua X, et al. Optimizing power to track brain degeneration in Alzheimer's disease and mild cognitive impairment with tensor-based morphometry: an ADNI study of 515 subjects. NeuroImage. 2009; 48:668–681. [PubMed: 19615450]

Jack C, et al. The Alzheimer's disease neuroimaging initiative (ADNI): the MR imaging protocol. Journal of Magnetic Resonance Imaging. 2008; 24:685–691.

Jebara, T. Machine learning: discriminative and generative. Boston: Kluwer; 2003.

Khan, A., Chung, M., Beg, M. Robust atlas-based brain segmentation using multi-structure confidence-weighted registration. Proceedings of international conference on medical image computing and computer assisted intervention (MICCAI'09), Part II; 2009. p. 549-557.

Klauschen F, Goldman A, Barra V, Meyer-Lindenberg A, Lundervold A. Evaluation of automated brain MR image segmentation and volumetry methods. Human Brain Mapping. 2009; 30:1310–1327. [PubMed: 18537111]

Klein A, et al. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. NeuroImage. 2009; 46(3):786–802. [PubMed: 19195496]

Lao, Z., Shen, D., Jawad, A., Karacali, B., Liu, D., Melhem, E., Bryan, N., Davatzikos, C. Automated segmentation of white matter lesions in 3D brain MR images, using multivariate pattern classification. Proceedings of IEEE international symposium on biomedical imaging (ISBI'06); Arlington. 2006. p. 307-310.

Lasserre, J., Bishop, C., Minka, T. Principled hybrids of generative and discriminative models. IEEE conference on computer vision and pattern recognition (CVPR'06); 2006. p. 87-94.

Lee, C., Schmidt, M., Murtha, A., Bistritz, A., Sander, J., Greiner, R. Segmenting brain tumor with conditional random fields and support vector machines. Proceedings of workshop of computer vision for biomedical image application: current techniques and future trends; Beijing. 2005. p. 469-478.

Leemput KV, Maes F, Vandermeulen D, Colchester A, Suetens P. Automated segmentation of multiple sclerosis lesions by model outlier detection. IEEE Transactions on Medical Imaging. 2001; 20(8): 677–688. [PubMed: 11513020]

Leemput KV, Maes F, Vandermeulen D, Suetens P. Automated model-based tissue classification of MR images of the brain. IEEE Transactions on Medical Imaging. 1999; 18(10):897–908. [PubMed: 10628949]

Li C, Goldgof D, Hall L. Knowledge-based classification and tissue labeling of MR images of human brain. IEEE Transactions on Medical Imaging. 1993; 12(4):740–750. [PubMed: 18218469]

Li, H., Fan, Y. Label propagation with robust initialization for brain tumor segmentation. IEEE international symposium on biomedical imaging (ISBI'12); 2012. p. 1715-1718.

Li Z, Fan J. 3D MRI brain image segmentation based on region restricted EM algorithm. Proceedings of SPIE. 2008; 6914:6914001–69,140.8.

Liang, P., Jordan, M. An asymptotic analysis of generative, discriminative, and pseudo-likelihood estimators. Proceedings of international conference on machine learning (ICML'08); 2008. p. 584-591.

Loncaric S. A survey of shape analysis techniques. Pattern Recognition. 1998; 31(8):983–1001.

Luo, Y., Chung, ACS. An atlas-based deep brain structure segmentation method: from coarse positioning to fine shaping. IEEE international conference on acoustics, speech and signal processing (ICASSP'11); 2011. p. 1085-1088.

Menze, BH., Leemput, KV., Lashkari, D., Weber, MA., Ayache, N., Golland, P. A generative model for brain tumor segmentation in multi-modal images. Proceedings of international conference on medical image computing and computer assisted intervention (MICCAI'10), Part II; 2010. p. 151-159.

Morra JH, Tu Z, Apostolova L, Green AE, Toga A, Thompson P. Comparison of adaboost and support vector machines for detecting Alzheimer'ss disease through automated hippocampal segmentation. IEEE Transactions on Medical Imaging. 2010; 29(1):30–43. [PubMed: 19457748]

Ng A, Jordan M. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. Advances in neural information processing systems (NIPS'02). 2002; 14:841–848.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Pham D, Prince J. Adaptive fuzzy segmentation of magnetic resonance images. IEEE Transactions on Medical Imaging. 1999; 18(9):737–752. [PubMed: 10571379]

Pieper, S., Lorensen, B., Schroeder, W., Kikinis, R. The NA-MIC kit: ITK, VTK, pipelines, grids and 3D slicer as an open platform for the medical image computing community. Proceedings of IEEE international symposium on biomedical imaging (ISBI'06); 2006. p. 698-701.

Pizer SM, et al. Deformable m-reps for 3D medical image segmentation. International Journal of Computer Vision. 2003; 55(2):85–106. [PubMed: 23825898]

Pohl K, Fisher J, Kikinis R, Grimson W, Wells W. A Bayesian model for joint segmentation and registration. NeuroImage. 2006; 31(1):228–239. [PubMed: 16466677]

Prastawa M, Bullitt E, Moon N, Leemput KV, Gerig G. Automatic brain tumor segmentation by subject specific modification of atlas priors. Academic Radiology. 2003; 10(12):134–1348.

Quddus, A., Fieguth, P., Basir, O. Adaboost and support vector machines for white matter lesion segmentation in MR images. IEEE annual conference on engineering in medicine and biology (EMBS'05); 2005.

Raina R, Shen Y, Ng A, McCallum A. Classification with hybrid generative/discriminative models. Advances in neural information processing systems (NIPS'03). 2003; 16:545–552.

Sabuncu M, Yeo B, Leemput KV, Fischl B, Golland P. A generative model for image segmentation based on label fusion. IEEE Transactions on Medical Imaging. 2010; 29(10):1714–1729. [PubMed: 20562040]

Scherrer B, Dojat M, Forbes F, Garbay C. LOCUS: local cooperative unified segmentation of MRI brain scans. Medical image computing and computer assisted intervention (MICCAI'07), Part I. 2007:219–227.

Scherrer B, Forbes F, Garbay C, Dojat M. Distributed local MRF models for tissue and structure brain segmentation. IEEE Transactions on Medical Imaging. 2009; 28:1278–1295. [PubMed: 19228553]

Schuff N, Tosun D, Insel P, Chiang G, Truran D, Aisen P, Jack CJ, Weiner MADN. Initiative: nonlinear time course of brain volume loss in cognitively normal and impaired elders. Neurobiology of Aging. 2012; 33:845–855. [PubMed: 20855131]

Shattuck D, Sandor-Leahy S, Schaper K, Rottenberg D, Leahy R. Magnetic resonance image tissue classification using a partial volume model. NeuroImage. 2001; 13:856–876. [PubMed: 11304082]

Shattuck DW, et al. Construction of a 3D probabilistic atlas of human cortical structures. NeuroImage. 2008; 39:1064–1080. [PubMed: 18037310]

Simpson, IJA., Woolrich, MW., Groves, AR., Schnabel, JA. Longitudinal brain MRI analysis with uncertain registration uncertain registration. Proceedings of international conference on medical image computing and computer assisted intervention (MICCAI'11), Part II; 2011. p. 647-654.

Sled J, Zijdenbos A, Evans A. A nonparametric method for automatic correction of intensity non-uniformity in MRI data. IEEE Transactions on Medical Imaging. 1998; 17:87–97. [PubMed: 9617910]

Smith S. Fast robust automated brain extraction. Human Brain Mapping. 2002; 17:143–155. [PubMed: 12391568]

Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. 1996; 58(1):267–288.

Tu, Z. Learning generative models via discriminative approaches. IEEE conference on computer vision and pattern recognition (CVPR'07); 2007. p. 1-8.

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

Tu Z, Narr K, Dollar P, Dinov I, Thompson P, Toga A. Brain anatomical structure segmentation by hybrid discriminative/generative models. IEEE Transactions on Medical Imaging. 2008; 27:495–508. [PubMed: 18390346]

Unay, D., Ekin, A., Jasinschi, R. Medical image search and retrieval using local binary patterns and KLT feature points. IEEE conference on image processing (ICIP'08); 2008. p. 997-1000.

Vapnik, V. Estimation of dependences based on empirical data. Berlin: Springer; 1982.

Vapnik, V. Statistical learning theory. New York: Wiley; 1998.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Verma R, Zacharaki EI, Ou Y, Cai H, Chawl S, Lee SK, Melhem ER, Wolf R, Davatzikos C. Multiparametric tissue characterization of brain neoplasms and their recurrence using pattern classification of MR images. Academic Radiology. 2008; 15(8):966–977. [PubMed: 18620117]

Wang H, Das S, Suh J, Altinay M, Pluta J, Craige C, Avants B, Yushkevich P. The Alzheimer's disease neuroimaging initiative: a learning-based wrapper method to correct systematic errors in automatic image segmentation: consistently improved performance in hippocampus, cortex and brain segmentation. NeuroImage. 2011; 55(3):968–985. [PubMed: 21237273]

Wells W, Kikinis R, Grimson W, Jolesz F. Adaptive segmentation of MRI data. IEEE Transactions on Medical Imaging. 1996; 15(4):429–442. [PubMed: 18215925]

Wels, M., Zheng, Y., Carneiro, G., Huber, M., Hornegger, J., Comaniciu, D. Fast and robust 3-D MRI brain structure segmentation. Proceedings of international conference on medical image computing and computer assisted intervention (MICCAI'09), Part II; 2009. p. 575-583.

Wels M, Zheng Y, Huber M, Hornegger J, Comaniciu D. A discriminative model-constrained EM approach to 3D MRI brain tissue classification and intensity non-uniformity correction. Physics in Medicine and Biology. 2011; 56(11):3269–3300. [PubMed: 21558592]

Wolz, R., Aljabar, P., Rueckert, D., Heckemann, R., Hammers, A. Segmentation of subcortical structures and the hippocampus in brain MRI using graph-cuts and subject–specific a-priori information. Proceedings of IEEE international symposium on biomedical imaging (ISBI'09); 2009. p. 470-473.

Woods R, Mazziotta J, Cherry S. MRI-PET registration with automated algorithm. Journal of Computer Assisted Tomography. 1993; 17:536–546. [PubMed: 8331222]

Woolrich M, Behrens T. Variational Bayes inference of spatial mixture models for segmentation. IEEE Transactions on Medical Imaging. 2006; 2(10):1380–1391.

Wu, J., Chung, A. A new solver for Markov random field modeling and applications to medical image segmentation. Proceedings of international conference on medical image computing and computer assisted intervention (MICCAI'05), Part I; 2005. p. 229-237.

Wu, J., Chung, A. Markov dependence tree-based segmentation of deep brain structures. Proceedings of international conference on medical image computing and computer assisted intervention (MICCAI'08), Part II; 2008. p. 1092-1100.

Wu J, Chung A. A novel framework for segmentation of deep brain structures based on Markov dependence tree. NeuroImage. 2009; 49:1027–1036.

Yang F, Shan ZY, Kruggel F. White matter lesion segmentation based on feature joint occurrence probability and χ^2 random field theory from magnetic resonance (MR) images. Pattern Recognition Letters. 2010; 31(9):781–790.

Yang J, Staib L, Duncan J. Neighbor-constrained segmentation with level set based 3D deformable models. IEEE Transactions on Medical Imaging. 2004; 23(8):940–948. [PubMed: 15338728]

Yaqub, M., Javaid, MK., Cooper, C., Noble, JA. Improving the classification accuracy of the classic rf method by intelligent feature selection and weighted voting of trees with application to medical image segmentation. International conference on machine learning in medical imaging (MLMI'11); 2011. p. 184-192.

Yi, Z., Criminisi, A., Shotton, J., Blake, A. Discriminative, semantic segmentation of brain tissue in MR images. Proceedings of international conference on medical image computing and computer assisted intervention (MICCAI'09), Part II; 2009. p. 558-565.

Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, Gerig G. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. NeuroImage. 2006; 31(3):1116–1128. [PubMed: 16545965]

Zhang, N., Ruan, S., Lebonvallet, S., Liao, Q., Zhu, Y. Multi-kernel SVM based classification for brain tumor segmentation of MRI multi-sequence. IEEE international conference on image processing (ICIP'09); 2009. p. 3373-3376.

Zhou J, Rajapakse J. Segmentation of subcortical brain structures using fuzzy templates. NeuroImage. 2005; 28(4):915–924. [PubMed: 16061401]

Zhu S, Yuille A. Region competition: unifying snake/balloon, region growing and Bayes/MDL/energy for multi-band image segmentation. IEEE Transactions on PAMI. 1996; 18(9):884–900.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

1: **Given**: A set of training volumes together with their image features $F(j)$ after applying $gmDT$, $S = \{(\mathbf{V}_n^{tr}, A_n^{tr}), n = 1..N\}$. The size of each volume is $|V|$.
2: For each volume V_n^{tr} , initialize probability maps $P_n^{(0)}$ with uniform distribution on all the labels.
3: Iteratively train T classifiers:
4: **for** $t = 1$ to T **do**
5: Make a training set of voxel samples from S: $\{(l_{ni}, (F(ni), P_n^{(t-1)}(i))), n = 1..N, i = 1..|V|\}$, where i is a voxel of volume V_n^{tr} and l_{ni} is from A_n^{tr}.
6: Train a classifier (in our implementation, a PBT classifier) on both image and context features from $F(ni)$ and $P_n^{(t-1)}(i)$ respectively. Output the classifier.
7: Use the trained classifier to compute new probability maps $P_n^{(t)}$ on all the labels for each training volume V_n^{tr}.
8: **end for**
9: The algorithm outputs a sequence of T trained classifiers for $p^{(T)}(l_i|F(i), P^{(T-1)}(i))$

Fig. 1.
Training process of auto-context algorithm for our brain segmentation method

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Given: N training volumes with their corresponding annotations $(V_1^{tr}, A_1^{tr}) \cdots ...(V_N^{tr}, A_N^{tr})$
Learn (V_a^{tr}, A_a^{tr}) from the N training volumes as the atlas.
for $i = 1$ to N **do**
  Obtain $\hat{W}_i^{tr} = \{\hat{W}_R, \hat{\Theta}\}$ by minimizing Equation 6. This step is called $gmDT$.
  for $j = 1$ to $|\Lambda|$ **do**
    Obtain normalized feature set $\mathbf{F_d}(i,j)$ based on $\hat{\Theta}$:
    $\left[f_{d,1}(\mathbf{V}_i(N_j)), \cdots, f_{d,B}(\mathbf{V}_i(N_j))\right]$
    $\rightarrow \left[f_{d,1}(\mathbf{V}_{i,\hat{\Theta}}(N_j)), \cdots, f_{d,B}(\mathbf{V}_{i,\hat{\Theta}}(N_j))\right]$
    Merge $\mathbf{F_d}(i,j)$ with the atlas features based on $\hat{W}_R$:
    $\mathbf{F}(i,j) \equiv \left[F_{\hat{W}_R}, f_{d,1}(\mathbf{V}_{i,\hat{\Theta}}(N_j)), \cdots, f_{d,B}(\mathbf{V}_{i,\hat{\Theta}}(N_j))\right]$
  end for
end for
Take samples from $\{A_i^{tr}(j), \mathbf{F}(i,j) : i = 1..n, j = 1..|\Lambda|\}$ to train a multi-class classifier $p(l|\mathbf{F})$.

Fig. 2.

Training process. All the training volumes are assumed to be skull-stripped and aligned

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

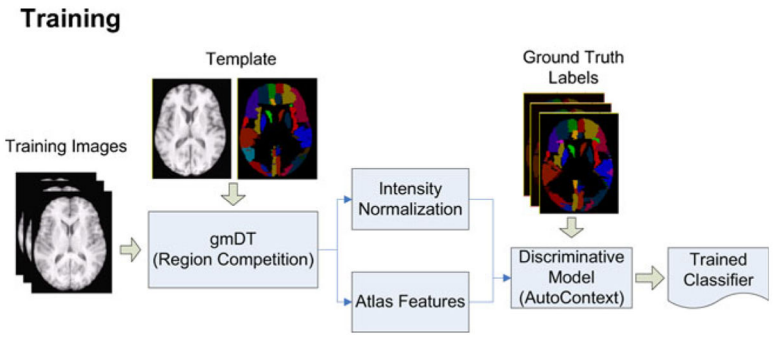


Fig. 3.

Training procedures of the proposed fusion method. A template of the volume/label pair is selected and guided by *gmDT* using other training volumes as the input. The augmented feature set ***F*** is then extracted to train the discriminative model giving the manual labels of the training volumes

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

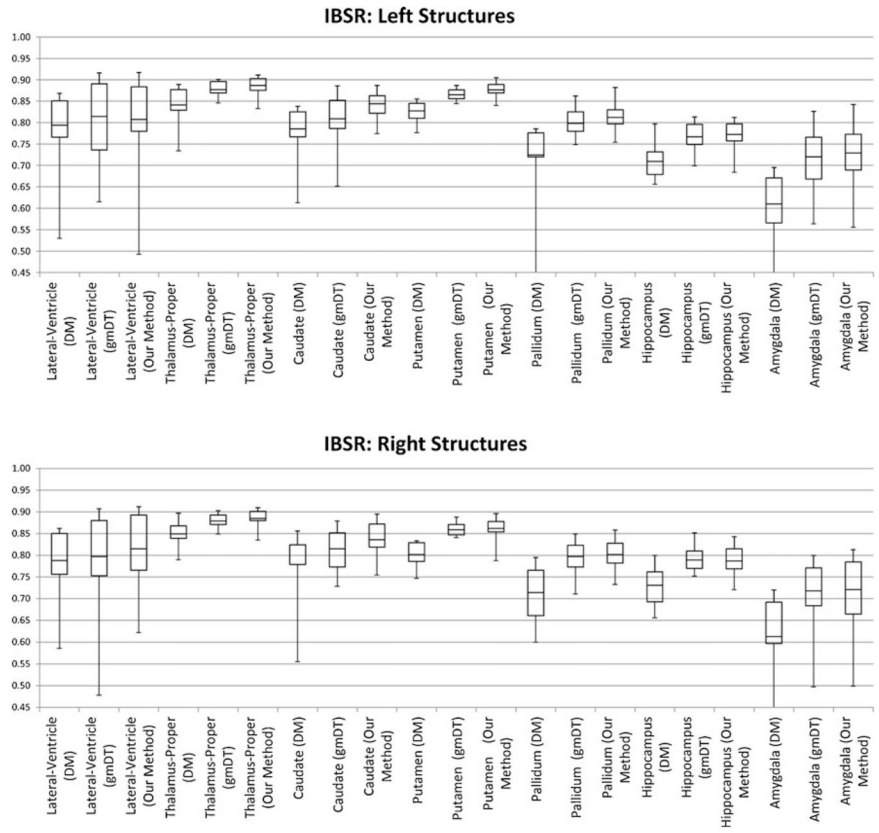


Fig. 4.

Box plots of the *Dice coefficients* of the discriminative model, the generative model based deformable template approach (*gmDT*), and our method on IBSR18. The top and bottom ends of a vertical line are the maximum and minimum values; The upper and lower edges of a box are the quartiles (25 % and 75 % data). The line inside a box indicates the average value, which is the same as Table 4

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

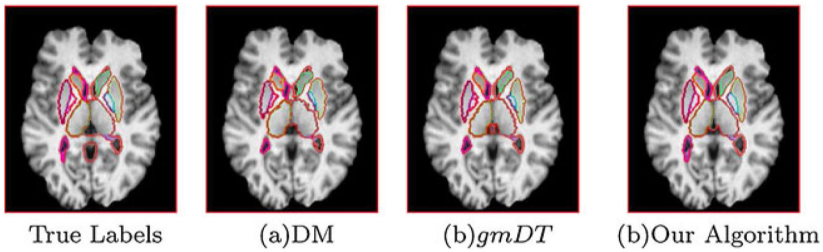


Fig. 5.

Segmentation results on a typical slice view by (**a**) discriminative method (DM), (**b**) the generative model based deformable template (*gmDT*), and (**c**) our algorithm. The leftmost picture shows the ground truth labeling for comparison. Only partial of the extracted structures are shown in this 2D view. The example image is IBSR_09 from the IBSR18 dataset

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

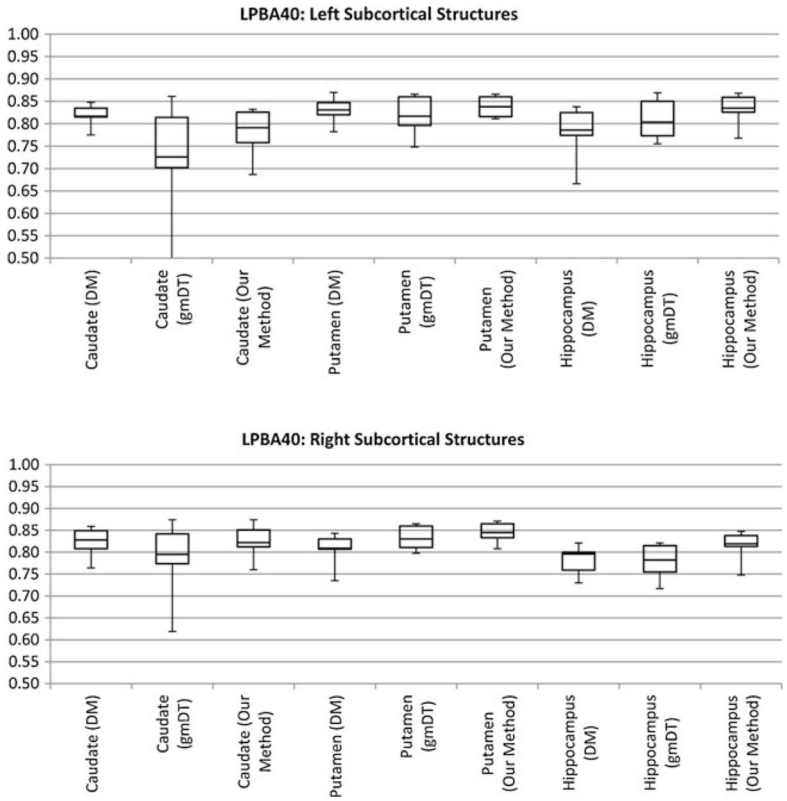
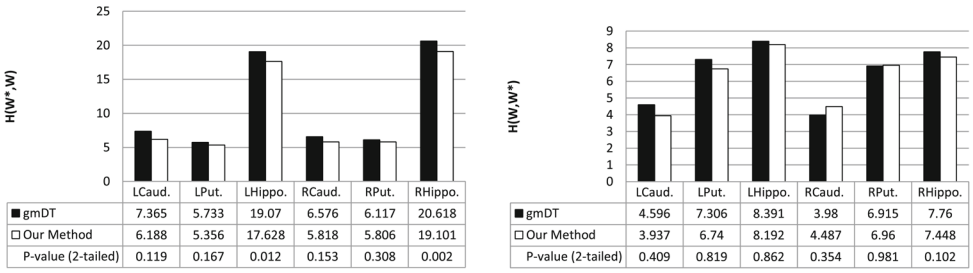


Fig. 6.

Box plots of the *Dice coefficients* of the discriminative model, the generative model based deformable template approach (*gmDT*), and our method on LPBA40. The top and bottom ends of a vertical line are the maximum and minimum values; The upper and lower edges of a box are the quartiles (25 % and 75 % data). The line inside a box indicates the average value, which is the same as Table 4. All models were trained by the 25 images from LPBA40



	LCaud.	LPut.	LHippo.	RCaud.	RPut.	RHippo.
■ gmDT	7.365	5.733	19.07	6.576	6.117	20.618
□ Our Method	6.188	5.356	17.628	5.818	5.806	19.101
P-value (2-tailed)	0.119	0.167	0.012	0.153	0.308	0.002

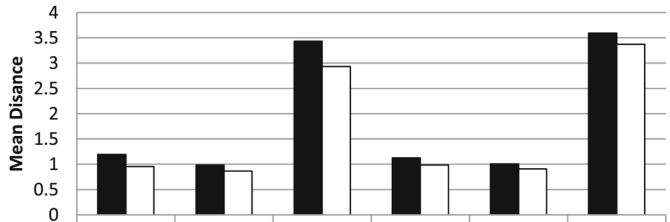
	LCaud.	LPut.	LHippo.	RCaud.	RPut.	RHippo.
■ gmDT	4.596	7.306	8.391	3.98	6.915	7.76
□ Our Method	3.937	6.74	8.192	4.487	6.96	7.448
P-value (2-tailed)	0.409	0.819	0.862	0.354	0.981	0.102

Fig. 7.

Inter-dataset *Hausdorff distance* measures in *mm*, on 40 LPBA40 volumes for extracting the three types of subcortical structures (smaller is better). We denote the automated segmented result as W^* and the ground truth as W.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript



	LCaud.	LPut.	LHippo.	RCaud.	RPut.	RHippo.
■ gmDT	1.192	0.982	3.43	1.126	1.004	3.593
□ Our Method	0.956	0.863	2.935	0.984	0.906	3.374
P-value (2-tailed)	0.001	0.004	0.0003	0.101	0.033	0.093

Fig. 8.

Inter-dataset *Mean distance* measures in *mm*, on 40 LPBA40 volumes for extracting the three types of subcortical structures (smaller is better). The standard deviations are shown in parentheses.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

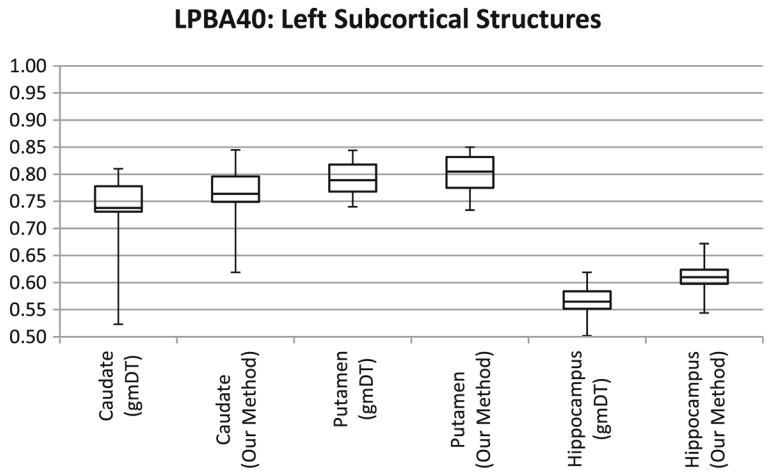
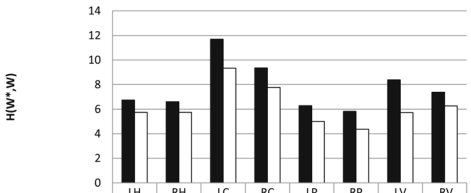
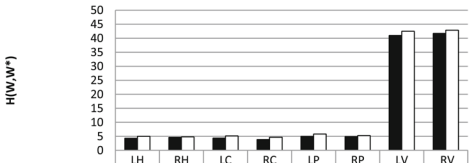


Fig. 9.

Box plots of the *Dice coefficients* of the generative model based deformable template approach (*gmDT*) and our method on LPBA40. The top and bottom ends of a vertical line are the maximum and minimum values; The upper and lower edges of a box are the quartiles (25% and 75% data). The line inside a box indicates the average value, which is the same as Table 5. The model was trained by IBSR18



	LH	RH	LC	RC	LP	RP	LV	RV
gmDT	6.743	6.606	11.699	9.36	6.284	5.818	8.388	7.374
Our Method	5.744	5.748	9.331	7.768	5	4.367	5.725	6.267
P-value (2-tailed)	0.005	0.006	0.021	0.034	0.001	3E-07	0.253	0.587



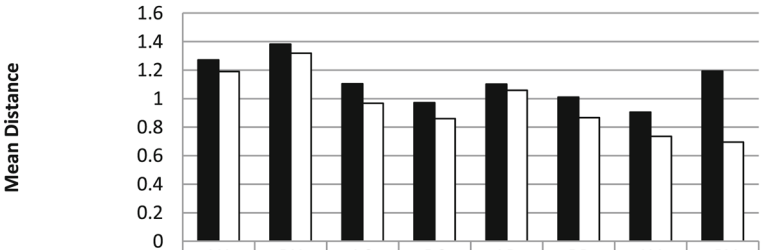
	LH	RH	LC	RC	LP	RP	LV	RV
gmDT	4.314	4.62	4.368	3.863	5	4.916	40.975	41.711
Our Method	4.993	4.811	5.124	4.659	5.837	5.276	42.502	42.862
P-value (2-tailed)	0.005	0.361	0.048	0.010	0.064	0.380	0.292	0.240

Fig. 10.

Inter-dataset *Hausdorff distance* measures in *mm*, on 28 LONI28 volumes for extracting the eight subcortical structures (smaller is better). We denote the automated segmented result as W^* and the ground truth as W

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript



	LH	RH	LC	RC	LP	RP	LV	RV
■ gmDT	1.271	1.382	1.105	0.972	1.102	1.011	0.906	1.193
□ Our Method	1.19	1.318	0.967	0.86	1.059	0.867	0.735	0.695
P-value (2-tailed)	0.200	0.315	0.063	0.031	0.494	0.0003	0.052	0.098

Fig. 11.

Inter-dataset *Mean distance* measures in *mm*, on 28 LONI28 volumes for extracting the eight subcortical structures (smaller is better). The standard deviations are shown in parentheses

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

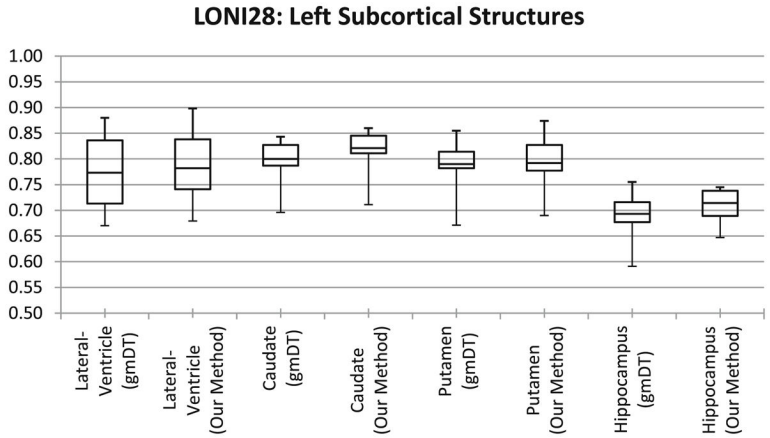
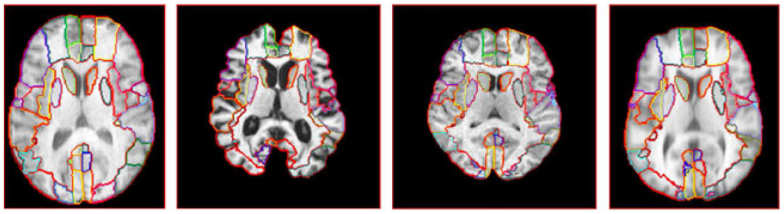


Fig. 12.

Box plots of the *Dice coefficients* of the generative model based deformable template approach (*gmDT*) and our method on LONI28. The top and bottom ends of a vertical line are the maximum and minimum values; the upper and lower edges of a box are the quartiles (25 % and 75 % ranked data). The line inside a box indicates the average value

Author Manuscript Author Manuscript Author Manuscript Author Manuscript



(a)LPBA40 (b)ADNI980 (c)LONI28 (d)IBSR18

Fig. 13.

Typical examples of the proposed brain segmentation method on different datasets. We use 2D slices of similar brain locations for comparison. Slice (a) is from LPBA40 to show the original imaging conditions of the training dataset. For clear comparison, these slices shown were skull-stripped and scaled to similar size. These four sets contain totally more than 1,000 volumes, and all the results are obtained by the identical system without parameter tuning

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Datasets used in this study. Note that we only use 14 subcortical structures in IBSR18 out of the 84 that are available

Dataset	Number of labels	Role in the study
IBSR18	84 (cortical+subcortical)	Training, Testing
LPBA40	56 (cortical+subcortical)	Training, Testing
LONI28	8 (subcortical)	Testing
ADNI980	0	Testing

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2

Distributions of the first 120 selected (most informative) features for a fixed annotation A_a and an adapted $\hat{W}_R$

	Atlas %	Haar %	Location %	Derivative %	Others %
Proportion	0.02	99.4	0.2	0.3	0.04
fixed atlas	2.9	48.1	31	18	0
adapted atlas	6	46	29	18	1

The first row shows the proportion of the corresponding type of features in the whole candidate feature pool for reference. The column “Others” includes the features of intensities, gradients, and curvatures

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 3

A comparison of our method and other methods on the IBSR18 dataset. We perform a three-fold cross-validation test on IBSR, each fold used 12 volumes for training and 6 volumes for testing

(a) Dice coefficients of *gmDT* and our method on the IBSR18 dataset

	LHippo.	**LCaud.**	**LPut.**	**LVent.**	**LPal.**	**LThal.**	**LAmyg.**
DM	0.709	0.785	0.827	0.794	0.724	0.841	0.610
	(0.036)	(0.056)	(0.021)	(0.082)	(0.085)	(0.036)	(0.065)
SyN	0.708	0.734	0.822	NA	0.732	0.838	0.650
	(0.021)	(0.045)	(0.008)	NA	(0.021)	(0.012)	(0.039)
gmDT	0.767	0.809	0.865	**0.814**	0.798	0.877	0.720
	(0.030)	(0.057)	(0.011)	(0.093)	(0.031)	(0.015)	(0.086)
Our Method	**0.772**	**0.844**	**0.876**	0.807	**0.812**	**0.887**	**0.729**
	(0.029)	(0.029)	(0.016)	(0.108)	(0.027)	(0.015)	(0.062)
	Rhippo.	RCaud.	RPut.	RVent.	RPal.	RThal.	RAmyg.
DM	0.731	0.777	0.802	0.788	0.714	0.849	0.613
	(0.040)	(0.072)	(0.028)	(0.073)	(0.057)	(0.024)	(0.104)
SyN	0.714	0.712	0.818	NA	0.732	0.840	0.647
	(0.021)	(0.053)	(0.012)	NA	(0.016)	(0.014)	(0.043)
gmDT	**0.789**	0.815	0.859	0.797	0.797	0.879	0.718
	(0.024)	(0.042)	(0.013)	(0.104)	(0.036)	(0.013)	(0.071)
Our Method	0.787	**0.836**	**0.862**	**0.815**	**0.802**	**0.885**	**0.721**
	(0.029)	(0.034)	(0.025)	(0.082)	(0.035)	(0.019)	(0.083)

(b) Related reports on IBSR subcortical structures (Khan et al. 2009)

Method	Hippo.	Caud.	Put.	Lateral-Vent.	Pal.	Thal.	Amyg.
Freesurfer	0.75	0.82	0.81	0.78	0.71	0.86	0.68
FSL	0.62	0.68	0.81	NA	0.73	0.85	0.59
(Khan et al. 2009)	0.76	0.83	0.87	0.85	0.72	**0.89**	0.66
(Wu and Chung 2009)	NA	0.80	0.82	NA	NA	0.85	NA
(Gouttard and et al. 2007)	0.69	0.76	0.78	0.84	0.72	NA	0.64
(Akselrod-Ballin et al. 2007)	0.69	0.80	0.79	NA	0.74	0.84	0.63
(Zhou and Rajapakse 2005)	0.70	0.80	0.81	NA	NA	0.84	0.64
(Wels et al. 2009)	0.73	0.80	0.82	NA	NA	NA	NA

Neuroinformatics. Author manuscript; available in PMC 2018 May 23.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(a) Dice coefficients of *gmDT* and our method on the IBSR18 dataset

	LHippo.	LCaud.	LPut.	LVent.	LPal.	LThal.	LAmyg.
(Luo and Chung 2011)	NA	0.78	0.80	NA	NA	0.84	NA
(Du et al. 2011)	0.72	0.80	0.83	**0.88**	0.77	0.83	**0.74**
Our Method	**0.78**	**0.84**	**0.87**	0.81	**0.81**	**0.89**	**0**.73

(a) The detailed Dice coefficients of *DM, gmDT*, and our method of the eight subcortical structures, where *DM* is the discriminative method and *gmDT* is the generative model based deformable template approach. The abbreviations are: L/RHippo.=Left/Right Hippocampus, L/RCaud.=Left/Right Caudate Nucleus, L/RPut.=Left/Right Putamen, L/RVent.=Left/Right Lateral Ventricle, L/RPal.=Left/Right Pallidum, L/RThal.=Left/Right Thalamus, L/RAmyg.=Left/Right Amygdala. The standard deviations of the Dice coefficients are shown in parentheses. (b) Other representative, subcortical segmentation results on the IBSR18 dataset (Khan et al. 2009). Values are the averages of the left and right structures if they were separately reported. The best values are marked in bold

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 4

Accuracy measures (*Dice coefficients*) on the LPBA40 test volumes for extracting the 56 structures

	LHippo.	LCaud.	LPut.	RHippo.	RCaud.	RPut.	Total 56 structures
DM	0.796	**0.813**	0.831	0.786	**0.828**	0.807	0.767(0.06)
gmDT	0.804	0.726	0.817	0.782	0.795	0.830	0.812(0.04)
Our Method	**0.835**	0.791	**0.838**	**0.819**	0.822	**0.845**	**0.822**(0.04)
p-value	0.017	0.136	0.053	0.003	0.051	0.063	2.5E-04

We show the average Dice of the 56 structures and the results of six subcortical structures which are highly referred in the later tests. The abbreviations are: L/RHippo.=Left/Right Hippocampus, L/RCaud.=Left/Right Caudate Nucleus, L/RPut.=Left/Right Putamen. Our model is trained on 25 volumes from LPBA40. Values in parentheses are the standard deviations and we mark the best measures in bold. DM is the discriminative method and *gmDT* is our generative model based deformable template approach. The two-sided p-values between our method and *gmDT* of the six structures are also shown in the last row

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 5

Inter-dataset measures (*Dice coefficients*) on 40 LPBA40 volumes for extracting the six subcortical structures

	Caudate	**Putamen**	**Hippocampus**
FreeSurfer	0.65(0.040)	0.64(0.105)	0.57(0.029)
FSL	0.63(0.063)	0.79(0.042)	0.53(0.043)
DM	0.73(0.058)	0.69(0.083)	0.58(0.037)
gmDT	0.74(0.064)	0.79(0.035)	0.57(0.029)
Our Method	**0.77**(0.045)	**0.80**(0.035)	**0.61**(0.029)

All values are the averages of the corresponding measures of the left and right structures, and the standard deviations are shown in parentheses. The model was trained from IBSR18. DM is the discriminative method and *gmDT* is our generative model based deformable template approach. The p-values between our method and *gmDT* are 1.58×10^{-5} (Caudate), 0.0135 (Putamen), and 4.87×10^{-14} (Hippocampus)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 6

Inter-dataset measures (*Dice coefficients*) on 28 LONI28 test volumes for extracting the eight subcortical structures

(a) FreeSurfer									
	LH %	**RH %**	**LC %**	**RC %**	**LP %**	**RP %**	**LV %**	**RV %**	**Av %**
Precision	47	54	78	78	71	77	81	72	70
Recall	66	84	77	77	83	83	76	76	78
F-value	0.55	0.66	0.77	0.77	0.77	0.80	0.78	0.74	0.74
(b) Hybrid (Tu et al. 2008)(discriminative+shape)									
	LH %	RH %	LC %	RC %	LP %	RP %	LV %	RV %	Av %
Precision	77	64	81	83	68	72	81	81	76
Recall	69	62	84	81	75	75	90	90	78
F-value	**0.73**	0.63	**0.82**	0.82	0.71	0.73	**0.85**	**0.85**	0.77
(c) *gmDT*									
	LH %	RH %	LC %	RC %	LP %	RP %	LV %	RV %	Av %
Precision	61	56	83	83	78	78	78	78	74
Recall	81	85	78	82	81	85	77	77	81
F-value	0.69	0.68	0.80	0.82	**0.79**	0.81	0.77	0.78	0.77
(d) Our Method									
	LH %	RH %	LC %	RC %	LP %	RP %	LV %	RV %	Av %
Precision	63	58	86	86	78	83	82	82	77
Recall	83	86	79	80	81	82	75	77	80
F-value	0.72	**0.69**	**0.82**	**0.83**	**0.79**	**0.83**	0.78	0.79	**0.78**

The measures of FreeSurfer and Hybrid tested on 14 volumes of LONI28 were from Tu et al. (2008), so we use the same measures (precision and recall rates) and F-values for comparison. Tu et al. (2008) can be considered as a combination of DM and a shape prior. *gmDT* is the generative model based deformable template approach. The abbreviations are: L/RH=Left/Right Hippocampus, L/RC=Left/Right Caudate Nucleus, L/RP=Left/Right Putamen, L/RV=Left/Right Lateral Ventricle, Av=Average. In Tu et al. (2008), the Hybrid model is trained by another 14 volumes of LONI28. Our model is trained from IBSR18 and still showed the best average values among the four methods. The best F-value of each structure is marked in bold

Neuroinformatics. Author manuscript; available in PMC 2018 May 23.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Neuroinformatics. Author manuscript; available in PMC 2018 May 23.

Table 7

12-month longitudinal analysis of the extracted hippocampi from the 490 subjects

(a) DM : DM fails to process the volumes from 13 subjects so the statistics here are based on 477 subjects. The difference between the three groups as well as the longitudinal change inside each group are demonstrated. However, the standard deviations of all measures are big.

	baseline average volume (mm^3)			**12-month average volume (mm^3)**		
	Left	**Right**	**Average**	**Left**	**Right**	**Average**
AD	2301.22	2372.67	2336.95	2009.52	2158.16	2083.84
MCI	2527.80	2717.52	2622.66	2342.51	2546.46	2444.48
Normal	3039.01	3082.53	3060.77	2927.31	2980.78	2954.05
	volume change (mm^3)			percentage loss (%)		
	Left	Right	Average	Left	Right	Average
AD	291.71(544)	214.52(506)	253.11(504)	10.23%(35%)	8.27%(30%)	9.44%(30%)
MCI	185.29(449)	171.06(504)	178.18(453)	7.30%(22%)	6.64%(24%)	7.12%(21%)
Normal	111.70(520)	101.75(592)	106.72(539)	1.65%(21%)	1.40%(24%)	1.67%(22%)

(b) *gmDT*: The extracted volumes are consistent due to the adaptation power of generative model, but it fails to represent the longitudinal changes of the three groups.

	baseline average volume (mm^3)			12-month average volume (mm^3)		
	Left	Right	Average	Left	Right	Average
AD	3545.03	3884.74	3714.89	3507.99	3778.29	3643.14
MCI	3607.94	3870.49	3739.21	3579.45	3846.53	3712.99
Normal	3751.34	4008.77	3880.05	3706.05	3962.01	3834.03
	volume change (mm^3)			percentage loss (%)		
	Left	Right	Average	Left	Right	Average
AD	37.04(413)	106.45(433)	71.75(382)	0.53%(13%)	2.18%(13%)	1.50%(11%)
MCI	28.49(408)	23.96(462)	26.22(401)	0.19%(11%)	0.21%(12%)	0.29%(11%)
Normal	45.28(359)	46.76(353)	46.02(345)	0.62%(9%)	0.64%(10%)	0.72%(9%)

(c) Our Algorithm: The results faithfully reflect both the absolute volumes and longitudinal changes of the three groups compared with DM (a) and *gmDT* (b).

	baseline average volume (mm^3)			12-month average volume (mm^3)		
	Left	Right	Average	Left	Right	Average
AD	2999.04	3115.78	3057.41	2837.53	2970.87	2904.20
MCI	3238.77	3415.67	3327.22	3352.16	3150.12	3251.14
Control	3788.27	3938.24	3863.26	3751.32	3897.77	3824.54
	volume change (mm^3)			percentage loss (%)		

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(a) DM : DM fails to process the volumes from 13 subjects so the statistics here are based on 477 subjects. The difference between the three groups as well as the longitudinal change inside each group are demonstrated. However, the standard deviations of all measures are big.

	baseline average volume (mm^3)			**12-month average volume (mm^3)**		
	Left	**Right**	**Average**	**Left**	**Right**	**Average**
	Left	Right	Average	Left	Right	Average
AD	161.52(161)	144.92(176)	153.22(141)	5.56%(6%)	4.96%(6%)	5.25%(5%)
MCI	88.65(168)	63.51(177)	76.08(149)	2.70%(6%)	1.79%(6%)	2.27%(5%)
Normal	36.95(153)	40.47(153)	38.71(136)	0.98%(4%)	1.06%(4%)	1.03%(4%)

DM is the discriminative model (Tu and Bai 2010) and *gmDT* is the generative model based deformable template approach. We demonstrate the changes in left hippocampus, right hippocampus, and the average of both sides. Individual volume change is obtained by the baseline volume subtracting the volume of 12-month after. Values listed are the averages of the 490 subjects, and those in parentheses are the standard deviations