A label fusion method using conditional random fields with higher-order potentials: Application to hippocampal segmentation,
Artificial Intelligence in Medicine, Volume 64, Issue 2, 2015, Pages 117-129, ISSN 0933-3657, https://doi.org/10.1016/j.artmed.2015.04.005.
(https://www.sciencedirect.com/science/article/pii/S0933365715000457)

A label fusion method using conditional random fields with higher-order potentials: Application to hippocampal segmentation

Carlos Platero, M.Carmen Tobar

Applied Bioengineering Group, Technical University of Madrid, Ronda de Valencia 3, 28012, Madrid, Spain.

Abstract

Objective: The objective of this study is to develop a probabilistic modeling framework for segmenting structures of interest from a collection of atlases. We present a label fusion method that is based on minimizing an energy function using graph-cut techniques.

Methods and materials: We use a conditional random field (CRF) model that allows us to efficiently incorporate shape, appearance and context information. This model is characterized by a pseudo-Boolean function defined on unary, pairwise and higher-order potentials. Given a subset of registered atlases in the target image for a particular region of interest (ROI), we first derive an appearance-shape model from these registered atlases. The unary potentials combine an appearance model based on multiple features with a label prior using a weighted voting method. The pairwise terms are defined from a Finsler metric that minimizes the surface of separation between voxels whose labels are different. The higher-order potentials used in our framework are based on the robust P^n model proposed by Kohli et al. The higher-order potentials enforce label consistency in cliques; hence, the proposed method can be viewed as an approach to integrate high-level information with images based on low-level features. To evaluate the performance and the robustness of the proposed label fusion method, we employ two available databases of T1-weighted (T1W) magnetic resonance (MR) images of human brains. We compare our approach with other label fusion methods in the automatic hippocampal segmentation from T1W-MR images.

Results: Our label fusion method yields mean Dice coefficients of 0.829 and 0.790 for the two databases used with mean times of approximately 80 and

160 seconds, respectively.
Conclusions: We introduce a new label fusion method based on a CRF model and on ROIs. The CRF model is characterized by a pseudo-Boolean function defined on unary, pairwise and higher-order potentials. The proposed Boolean function is representable by graphs. A globally optimal binary labeling is found using a st-mincut algorithm in each ROI. We show that the proposed approach is very competitive with respect to recently reported methods.

Keywords:
Atlas-based segmentation, Image registration, Label fusion, Graph cuts, Global optimization, Hippocampal segmentation, Magnetic resonance imaging

---

1. Introduction

The automatic segmentation of subcortical structures in human brain magnetic resonance (MR) images plays a crucial role in clinical practice. The extraction of biomarkers from MR images is directed at variations in subcortical shapes and their volume measurements [1, 2]. This task is very important but difficult to perform, even by hand. Neuroanatomists often develop and use complicated protocols in guiding the manual delineation process [3, 4]. Specifically, hippocampal segmentation is an important tool for studying neurodegenerative diseases. Hippocampal volume and shape measures are commonly used as biomarkers for Alzheimer's disease, epilepsy and schizophrenia, among other [5]. This structure is difficult to segment because of its small size, high variability and low contrast in MR images.

Many approaches have been proposed, and most segmentation methods use deformable surfaces or atlas-based techniques. Deformable methods tend to explicitly use learned shape variation as a priori information in segmentation. Various methods for representing shapes and their relationships with the appearances have been proposed, such as region growing [6], level-set within a Bayesian framework [7], probabilistic boosting tree [8] or using active shape-appearance models [9]. Atlas-based segmentation has become a standard technique for identifying structures from brain MR images. The atlas-based methods have been demonstrated to outperform other algorithms [10]. The atlas-based approaches are generally based on non-rigid registrations. In the context of this study, an atlas is an image in one modality with its

respective labeling (typically generated by manual segmentation) [11]. The atlas-based methods allow a priori knowledge about the appearance and the shape of the anatomical structures to be introduced in a relatively simple way: only a registration method and a number of pre-segmented data sets are required. Barnes et al. [12] proposed registering the most similar atlas from an atlas set to segment the hippocampus. However, segmentations with a single atlas are intrinsically biased toward the shape and the appearance of a subject. Several studies have shown that approaches that incorporate the properties of a group of atlases outperform those that use a single atlas [11, 13, 14, 15, 16]. The primary benefit of the multi-atlas segmentation approach is that the effect of the errors associated with any single atlas propagation can be reduced in the process of combination. The transferred atlases are used to construct a model for segmenting the target image. This process is often called label fusion. Therefore, there are two steps in the multi-atlas based segmentation: (1) image registration and (2) label fusion. We focus on label fusion in this paper.

The label fusion methods have been classified into two categories: global weighted voting and local weighted voting. Most existing label fusion methods are based on global weighted voting, such as majority voting (MV) [13], STAPLE [17] and weighted voting (WV) [18], which are widely used in medical image segmentation. In these approaches, each atlas contributes to the resulting segmentation with the same weight for all of its voxels. It is very sensitive to the registration errors because it does not take into account the relevance of each sample. Recent works have shown that local weighted voting methods outperform global weighted voting methods [18, 19, 20].

Two approaches can be used to take into account the information of each voxel: (i) the atlases are registered non-rigidly into the target image [11, 15, 16] and (ii) the atlases are aligned in the target image and a patch-based label fusion method is applied [20, 21, 22, 23, 24]. The first approach has the advantage of forcing the resulting segmentation to have a similar global shape to those of expert-labeled structures in the atlases. There is a one-to-one mapping between the target image and each atlas. The label fusion methods of this approach generally calculate the labeling associated with the target image via maximum a posteriori (MAP) estimation [15, 19, 25]. In contrast to fusing label maps using non-rigid registrations, the second approach is based on the nonlocal mean principle [26]. This second approach increases the number of samples considered during the labeling estimation. The typical assumption of one-to-one mapping in non-rigid registration-based

techniques is relaxed through the use of local search windows. However, the labeling is local and independent, without global constraints. In this article, we focus on the label fusion methods that use non-rigid registrations. We leave the patch-based labeling methods for future studies.

Approach

The new segmentation algorithms attempt to integrate high-level information with image-based low-level features. At the low level, the appearance of an image patch leads to ambiguities in its labels. For example, in the case of the hippocampal head, the appearance of this structure is convoluted and blends with the amygdala. To overcome these ambiguities, it is necessary to incorporate extra information, such as a priori shape information and contextual information. In medical images, context plays a very important role because the anatomical structures are mostly constrained to relatively fixed positions. From the Bayesian perspective, context information is carried in the joint multivariate statistics in the posterior probability, which is often decomposed into likelihood and prior. In image processing, likelihood and prior often correspond to appearance and shape, respectively.

Brain images present different structures of interest to be segmented. A region-wise approach is more appropriate [27], which can be achieved by dividing the image into multiple anatomically meaningful regions [28]. Therefore, our task is to segment a given 3D target image into K anatomical structures, where K is fixed. We assume that it is possible to define a region of interest (ROI) such that its voxels only belong to a k-structure or to the background. Partitioning the problem into ROIs improves the results of registrations and segmentations. Indeed, the multi-atlas approaches have greater accuracy when the registrations are only made near the object of interest and not in the entire image [28]. Furthermore, these approaches convert the complex multi-label problem into feasible binary segmentation problems. For each ROI, a segmentation is denoted as S_k, and the optimal solution S_k^* can be obtained by the following Bayesian framework:

$$S_k^* = \arg\max_{S_k} p(I_k|S_k;\Theta_k)p(S_k;\Theta_k)$$

where I_k is the target image in the k-ROI and $p(I_k|S_k;\Theta_k)$ and $p(S_k;\Theta_k)$ define the image likelihood and the shape prior of the ROI, respectively. The classification model parameters for the k-ROI are denoted by Θ_k. In general, either a generative or discriminative model is used for the image likelihood,

whereas the shape models use the transferred labels from the registered atlases combined with simple geometric constraints. In terms of appearance, generative models have explicit model parameters and are able to capture the global variability, but they often have simplified assumptions, which limits their ability to model inhomogeneous patterns. By contrast, in a discriminative appearance model, there are no explicit parameters to estimate. Discriminative models are able to combine many of the local statistics, which are insensitive to complex and inhomogeneous texture patterns [8]. However, these models have difficulty in taking the regional information into account. For this reason, discriminative appearance models are combined with shape prior models [8, 29, 30].

Considerable research effort has been devoted to developing efficient algorithms for estimating the MAP solution. The simplest approach is to treat the segmentation as independent voxels and apply the standard classification algorithms. This approach is straightforward, but it loses the important interdependency information. The other extreme of the solution is to treat each instance of L as a single label and estimate its posterior probability. This implementation is infeasible because the space of the output labels grows exponentially with the size of the image. Conditional random fields (CRF) [31] have been widely used to model the correlations of the structured labels. The use of a CRF allows appearance, shape and context to be incorporated in a single unified model, although there are also other alternative approaches [29, 32, 33].

However, CRF models are typically defined on the basis of a fixed neighborhood structure and make unrealistic conditional independence assumptions, thereby limiting their modeling capabilities. Segmentations using only unary-pairwise potentials tend to over-smooth, and they also have difficulties in capturing global shapes. To overcome these drawbacks, CRF models can be improved through the use of higher-order potentials defined on sets of voxels or cliques [34]. In this study, a CRF model is used for fusing the registered atlases, and this model is characterized by an energy function defined on unary, pairwise and higher-order potentials. The unary potentials of the CRF model are defined as the negative log of the likelihood of a label being assigned to a voxel. It is computed from an appearance model and a label prior. The pairwise edge potentials have the form of a spatial regularizer that minimizes the surface of separation between two different labels [35]. The conventional unary and pairwise cues are coupled with higher-order potentials that are defined on voxel sets generated using *textons* (a texton is a

label given to a voxel that describes the local texture and associated through an appearance vector [36, 37]). The higher-order potentials enforce label consistency in image regions, and the proposed method can be considered an approach to integrate high-level information with image-based low-level features.

Although the experiments presented in this paper focus on hippocampal segmentation, the proposed concepts are generic and could be incorporated into other modalities and applications. We test different label fusion methods on publicly available MR images of human brains. We show that our approach produces segmentation results that are as good as or better than those of other label fusion methods.

The remainder of this paper is organized as follows. In Section 2, the label fusion method is presented. The hippocampal segmentation experiments are described in Section 3. Finally, the discussion and conclusions are presented in Section 4.

2. Label fusion method

We present a label fusion method that is based on minimizing a pseudo-Boolean function using graph cuts with information on appearance, shape and context, which are estimated from the registered atlases in the target image. Other authors have previously used this framework [15, 25, 38, 39]. Our label fusion method has the following differences: a) a generative/discriminative appearance model based on multiple features extracted from each voxel and its neighborhood, b) a label prior probability is estimated using a weighted voting method [18], c) a spatial regularizer that minimizes the surface of separation between two different labels [35], and d) higher-order potentials are used to obtain label consistency in the cliques.

Given a ROI in the target image I, a set of N training atlases $\{A_i\}_{i=1,\ldots,N} = \{I_i, S_i\}_{i=1,\ldots,N}$ are used for the label fusion method, where $I_i : \Omega_i \subset \mathbb{N}^n \rightarrow \mathbb{R}$, $n = 3$, are the modality images and $S_i : \Omega_i \subset \mathbb{N}^n \rightarrow \{0, 1\}$ are the label maps. In the labeled images, voxels that belong to the k-structure are designated by the label $S(x) = 1$, and background voxels are designated by the label $S(x) = 0$. We denote $\Phi_i : \Omega \rightarrow \Omega_i$ to be the spatial mapping from the target image coordinates to the coordinates of the $i-$th atlas. For simplicity, we assume that $\{\Phi_i\}_{i=1,\ldots,N}$ was pre-computed using a pairwise registration procedure. This assumption allows us to simplify

$\mathbb{A} = \{\tilde{S}_i = S_i \circ \Phi_i, \tilde{I}_i = I_i \circ \Phi_i\}_{i=1,\ldots,N}$ as the atlases in the coordinates of the target image. We seek to minimize an energy function under the Bayesian formulation, which defines the conditional probability as a discrete random field S with a neighborhood system $\mathcal{E}$ and a clique set $\{C_b\}_{b=1\ldots,B}$, where a clique $C_b \subset \Omega$ is a connected set of voxels whose labels are conditionally dependent on each other. The neighborhood system $\mathcal{E}$ is the set of edges that connect variables in the random field. The CRF model is defined by the following pseudo-Boolean function:

$$\begin{aligned} E(S) = & \sum_{x \in \Omega} \psi_x(S(x); \theta_1(I, \mathbb{A})) \\ & + \sum_{x,y \in \mathcal{E}} \psi_{xy}(S(x), S(y); \theta_2(I)) \\ & + \sum_{b=1}^{B} \psi_{C_b}(S(C_b); \theta_C(I, \mathbb{A})) \end{aligned} \tag{1}$$

where $\Theta = \{\theta_1, \theta_2, \theta_C\}$ are the model parameters for this ROI. We next define the form of the three potential functions and their parameters.

2.1. The unary potentials

The unary potentials $\psi_x(S(x); \theta_1(I, \mathbb{A}))$ use the Bayesian formulation, which allows a priori information about the shape and appearance of structures to be segmented to be incorporated. As we experimentally demonstrate, unary potentials are the most powerful terms in the CRF model.

Image likelihood

We assume that the observed intensities of I are independent random variables. The image likelihood $p(I|S; \mathbb{A})$ can then be written as a product of the likelihoods of the individual voxels:

$$p(I|S; \mathbb{A}) = \prod_{x \in \Omega} p(I(x)|S(x); \mathbb{A}).$$

The following two approaches are compared: a) a generative appearance model in which a Gaussian quadratic classifier is defined for each voxel, and b) a discriminative appearance model based on k-nearest neighbor voting. We start by developing the generative model.

In general, the intensity distribution is modeled using a mixture of Gaussians [40, 41]. Because there is a one-to-one mapping between the target image and each atlas, we alternatively use a multivariate Gaussian distribution for each voxel and for each label [27, 42]. A pool of feature candidates are extracted from the training images, such as intensity, gradients, curvatures, and entropies. A feature selection process is applied, and $G_I(x)$ denotes the selected feature vector centered at x from I (for further details, see section 3.1.1). The predicted appearance of a voxel is defined by

$$p(I(x)|l;\mathbb{A}) \propto \frac{1}{|\Sigma_l(x)|^{1/2}} \cdot e^{\left(-\frac{1}{2}(G_I(x)-\mu_l(x))^T\Sigma_l^{-1}(x)(G_I(x)-\mu_l(x))\right)}, \tag{2}$$

where $l \in \{0,1\}$, μ is the mean vector, and Σ is the covariance matrix. The effect of sample size on the feature selection has to be considered. The means and covariance matrices are estimated using a variable number of samples $\#Q_l(x)$, where $Q_l(x) = \{i|\tilde{S}_i(x) = l\}$. A minimum number of observations is required from each of the two classes to ensure that the classification error is bounded relative to an infinite number of samples. This number depends on the dimension of the feature space. Let d be the dimension of the feature space. We have $d \leq \frac{1}{5}\min(\#Q_0(x), \#Q_1(x))$ for $d \leq 8$ [43]. To obtain the least biased Gaussian parameters, a neighborhood system around the voxel is used to obtain more samples, $\mathcal{N}(x)$. The Gaussian parameters are computed from $\mathbb{A}$:

$$\mu_l(x) = \frac{\sum_{y\in\mathcal{N}(x)}\sum_{i\in Q_l(y)} G_{\tilde{I}_i}(y)}{\sum_{y\in\mathcal{N}(x)}\#Q_l(y)} \tag{3}$$

and

$$\Sigma_l(x) = \frac{\sum_{y\in\mathcal{N}(x)}\sum_{i\in Q_l(y)}(G_{\tilde{I}_i}(y)-\mu_l(x))(G_{\tilde{I}_i}(y)-\mu_l(x))^T}{\sum_{y\in\mathcal{N}(x)}\#Q_l(y)-1}. \tag{4}$$

Furthermore, d is variable in each voxel. The correlation matrix over the selected features is analyzed for each voxel. It only selects uncorrelated features during runtime.

Then, a discriminative appearance model that requires low computational effort is presented. The registered atlas images are convolved using a filterbank. A set of feature extraction kernels α_j (for further details, see section 3.1.1) is used to produce different feature maps:

$$F_{\tilde{I}_i}(x) = \{\tilde{I}_i(x) * \alpha_j(x)\}_{j=1,\ldots,f}$$

where f is the dimension of this feature vector and $F_{\tilde{I}_i}(x)$ denotes the resulting feature vector of $\tilde{I}_i$ at voxel x associated with the filter-bank $\{\alpha_j(x)\}_{j=1,\ldots,f}$. In this paper, derivatives of Gaussians are adopted to extract features [36, 37]. The responses for all registered atlas image voxels are whitened separately (to provide zero mean and unit covariance). These feature vectors are used to train a k-nearest neighbor (k-NN) appearance model. For computational efficiency, we use the kd-tree algorithm [44] to perform the nearest neighbor search. A kd-tree model is constructed with $\{F_{\tilde{I}_i}(x), \tilde{S}_i(x)\}_{i=1,\ldots,N, x\in\Omega}$. The target image is also convolved and whitened. Let $\mathcal{F}_l = \{F_{\tilde{I}_i}(x) \,/\, \tilde{S}_i(x) = l\}$ be the set of feature vectors extracted from the voxels belonging to the registered atlas images and whose labels are l. Let $\{F_r\}_{r\in R_l(x)} \subset \mathcal{F}_l$ be the set whose elements are nearest neighbors to $F_I(x)$ and $R_l(x)$ be the set of the indices of the feature vectors with label l that are nearest neighbors to $F_I(x)$. The image likelihoods of the individual voxels belonging to the target image are calculated using the following formula:

$$p(I(x)|l; \mathbb{A}) \propto \sum_{r\in R_l(x)} \exp\left(-\|F_I(x) - F_r\|_2^2\right) \tag{5}$$

Label prior

The label prior probability $p(S; \mathbb{A}, I)$ models the joint probability of all voxels that belong to the ROI in a particular label configuration. Instead, we assume that the prior probability that voxel x has label l only depends on its position, the similarity between I and $\tilde{I}_i$ and the transferred atlas labeled images:

$$p(S; I, \mathbb{A}) = \prod_{x\in\Omega} p(S(x); I, \mathbb{A}).$$

This assumption is not realistic, but we encode the correlations of the labels using pairwise and higher-order potentials. For each voxel x and each label $l \in \{0, 1\}$, we define:

$$h(S(x) = l; I, \mathbb{A}) = \sum_{i\in Q_l(x)} m(I, \tilde{I}_i, x)^q$$

where $m(I, \tilde{I}_i, x)$ is a local or global similarity measure between the target image and the registered atlas image at x and q is an associated gain exponent [18]. The prior probability is defined as

$$p(S(x) = l; I, \mathbb{A}) = \frac{h(S(x) = l; I, \mathbb{A})}{\sum_{j \in \{0,1\}} h(S(x) = j; I, \mathbb{A})}. \tag{6}$$

Image likelihood and label prior terms are combined to define the unary potentials $\psi_x(S(x); \theta_1(I, \mathbb{A}))$:

$$\psi_x(S(x); \theta_1(I, \mathbb{A})) = -\log\left(\frac{p(I(x)|S(x); \mathbb{A})p(S(x); I, \mathbb{A})}{p(I(x; \mathbb{A}))}\right).$$

2.2. Spatial regularization

Following the work of Boykov and Kolmogorov [35], a smoothness term is added to the energy function. These authors decomposed this term, which is defined from a Finsler metric, into two elements. The first part minimizes the segmentation surface by a Riemannian metric, and the second one takes into account the orientation of the segmentation surface in the metric. We consider two types of Riemannian metrics from the image: a) $D(x) = g(\|\nabla I(x)\|)\mathbb{I}$, which is an isotropic metric, and b) $D(x) = g(\|\nabla I(x)\|)\mathbb{I} + (1 - g(\|\nabla I(x)\|)) \cdot u \cdot u^T$, which is anisotropic, where $\mathbb{I}$ is the identity matrix of size n, $g(\|\nabla I(x)\|) = (\exp(-\|\nabla I(x)\|/\gamma))^{1/3}$, γ is estimated by the average of $\|\nabla I(x)\|$, and $u = \frac{\nabla I(x)}{\|\nabla I(x)\|}$. The cubic root for the term $g()$ is a mapping between the values in $\|\nabla I(x)\|$ and $g()$. These pairwise potentials take the form of a contrast-sensitive Potts model,

$$\psi^R_{xy}(S(x), S(y); \theta_2(I)) = \begin{cases} 0 & \text{if } S(x) = S(y), \\ \frac{\det D(x)}{(\vec{xy}^T \cdot D(x) \cdot \vec{xy})^2} & \text{otherwise.} \end{cases}$$

The second term uses the flux of a vector field $\vec{v}(x)$ through the segmentation surface. It is assumed that the neighborhood system $\mathcal{E}$ of the CRF model is symmetric, that the vector $\vec{v}(x)$ can be decomposed into the set of edges belonging to $\mathcal{E}$ and that there are no local changes in the vector field. Under these constraints, the flux of the vector field defines the following pairwise potentials:

$$\psi^f_{xy}(S(x), S(y); \theta_2(I)) = \begin{cases} 0 & \text{if } S(x) = S(y), \\ \vec{v}(x) \cdot \frac{\vec{xy}}{\|\vec{xy}\|^2} & \text{if } S(x) = 0,\ S(y) = 1, \\ -\vec{v}(x) \cdot \frac{\vec{xy}}{\|\vec{xy}\|^2} & \text{if } S(x) = 1,\ S(y) = 0. \end{cases}$$

2.3. Higher-order potentials

A new higher-order potential is proposed based on the robust P^n model [34], which enforces label consistency softly. This potential uses a clique set from I, which is obtained using an unsupervised segmentation algorithm based on textons [36, 37]. The target image is first labeled in textons, and then the cliques are formed by spatially connecting voxels that have equal textons. Not all cliques are equally good; some cliques contain voxels with different labels. The probabilities previously calculated for the unary potentials are used to define the label consistency of each clique. Let $C = \{x_1, \ldots, x_s | x_i \in \Omega\}$ be a clique belonging to I. We define the probability that a clique belongs to a label l as

$$p(S_l(C); I, \mathbb{A}) = \frac{1}{\#C} \sum_{x_i \in C} p(S(x_i) = l | I(x_i); I, \mathbb{A}).$$

The non-dominant label of a clique and its probability are inferred from:

$$\begin{aligned} l_{min} &= \arg\min_l \left(p(S_l(C); I, \mathbb{A}) \right), \\ p_{l_{min}} &= p(S_{l_{min}}(C); I, \mathbb{A}). \end{aligned} \tag{7}$$

Assuming that all of the voxels belonging to a clique are equiprobable to obtain l_{min}, a new potential is defined as:

$$\psi_C \left(S(C); \theta_C(I, \mathbb{A}) \right) = \begin{cases} 0 & \text{if } S(x_i) = 1 - l_{min},\ \forall x_i \in C, \\ n_{l_{min}} & \text{if } n_{l_{min}} \leq p_{l_{min}} \#C, \\ \gamma_{max} & \text{if } n_{l_{min}} \geq p_{l_{min}} \#C, \end{cases} \tag{8}$$

where $n_{l_{min}}$ is the number of variables in the clique C whose labels are l_{min} and $\gamma_{max} = p_{l_{min}} \#C$. This potential increases linearly with $n_{l_{min}}$ up to γ_{max} as the robust P^n model [34]. However, our potential should be approximated to a step function around $p_{l_{min}} \#C$, i.e., if $n_{l_{min}} < p_{l_{min}} \#C$, then the label consistency cost is close to 0; otherwise, it is γ_{max} (see Fig. 1). The issue is

that this potential is not representable by graphs. To overcome this drawback, the voxels are weighted to preserve the consistency of the labeling of the clique. Let $\omega_i^l \geq 0$ be a weight that is used to specify the relative importance of label l in x_i. Then, these weights are normalized as $\omega_i^l \leftarrow \frac{\omega_i^l \cdot \#C}{\sum_i \omega_i^l}$. This step allows an average voxel to not lose its unitary cost if the label changes.

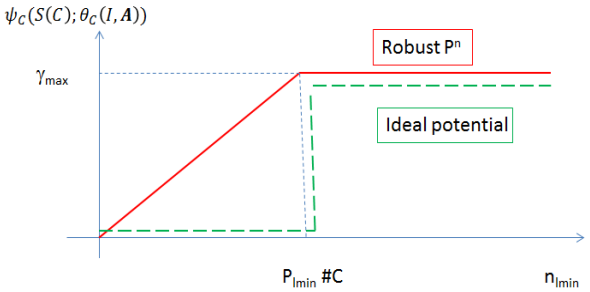

Figure 1: Behavior of the higher-order potentials with the number of voxels whose labels are not dominant

Now, $n_{l_{min}}$ is replaced by $\sum_i \omega_i^{l_{min}} \delta_{l_{min}}(S(x_i))$, where $\delta_l(S(x_i))$ is the Kronecker delta function that returns 1 if $S(x_i) = l$ and 0 otherwise. Inserting the weight of each voxel on the consistency of the labeling in (8), the family of higher-order potentials can be written as

$$\psi_C\left(S(C);\theta_C(I,\mathbb{A})\right) = \min\left\{\sum_i \omega_i^{l_{min}} \delta_{l_{min}}(S(x_i)), \gamma_{max}\right\}.$$

This potential family can be viewed as a weighted version of (8). The weights can be used to specify the relative importance of different voxels. These higher-order potentials can be transformed to pairwise potentials through the addition of one auxiliary binary variable (see Fig. 2). In the case of $l_{min} = 0$, we have

$$\psi_C\left(S(C);\theta_C(I,\mathbb{A})\right) = \min_{z\in\{0,1\}}\left\{\gamma_{max} + \left(-\gamma_{max} + \sum_i \omega_i^0 \delta_0(S(x_i))\right) z\right\}.$$

If $l_{min} = 1$, then

$$\psi_C\left(S(C);\theta_C(I,\mathbb{A})\right) = \min_{z\in\{0,1\}}\left\{\gamma_{max} + \left(-\gamma_{max} + \sum_i \omega_i^1 \delta_1(S(x_i))\right)(1-z)\right\}.$$

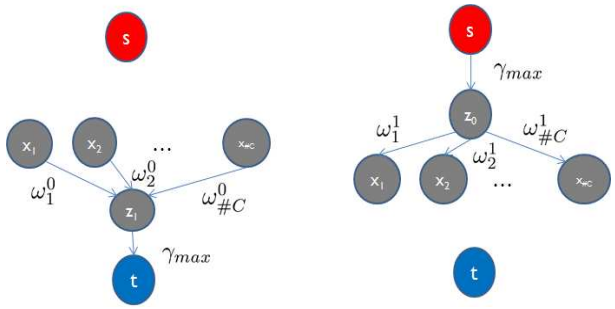

Figure 2: Graphs representing the proposed higher-order potentials: a) $l_{min} = 0$ and b) $l_{min} = 1$

We also define the quality of the clique by the formula

$$G(C) = \exp\left(-\frac{\sum_{x_i \in C} \|F_I(x_i) - \Upsilon\|_2^2}{\beta \cdot \#C}\right)$$

where $F_I(x_i)$ is the feature vector for labeling a voxel with a texton, β is an outlier measure of the dispersion of $\|F_I(x_i) - \Upsilon\|_2^2$ for all cliques (i.e., $\{C_b\}_{b=1...,B}$), and Υ is the centroid of the texton that is assigned to the clique C. The function $G(C)$ provides a measure of the compactness of the clique C. Consequently, γ_{max} and ω_i^l (which was normalized) have to be modified as

$$\gamma_{max} = p_{l_{min}} \#C \cdot G(C)$$
$$\omega_i^l \leftarrow \omega_i^l \cdot G(C).$$

A low level of $G(C)$ indicates that the clique has little consistency and that this potential does not have an influence on the labeling of the voxels in C. Conversely, if $G(C)$ is close to one, the voxels of the clique have a very similar appearance and these higher-order potentials will force a voxel subset of the clique to be labeled with the non-dominant label in relation to $p_{l_{min}}$.

The proposed higher-order potential family only affects the cliques whose voxels have the same appearance. Precisely classifying the voxels in these cliques is more difficult. In these cases, the higher-order potentials use $p(S_{l_{min}}(C); I, \mathbb{A})$ and the weights $\omega_i^{l_{min}}$. For example, Fig. 2.a shows the proposed graph when the non-dominant label in a clique is '0'. For the voxels whose weights tend to be smaller ($\omega_i^0 \to 0$), these potentials will favor their

labelings as '0' in relation to $p(S_0(C); I, \mathbb{A})$. A clique with noise or that is not homogeneous in appearance makes $G(C)$ tend to 0; thus, its higher-order potential does not affect its labeling.

2.4. The proposed CRF model

Introducing the three potential functions in (1), the weighting multipliers $\Lambda = \{\lambda_R, \lambda_f, \lambda_C\}$ are tuned such that the effects of the different model potentials are recombined to obtain the best segmentation results:

$$E(S) = \sum_{x \in \Omega} \psi_x(S(x); \theta_1(I, \mathbb{A})) + \lambda_R \sum_{x,y \in \mathcal{E}} \psi^R_{xy}(S(x), S(y); \theta^R_2(I)) + \lambda_f \sum_{x,y \in \mathcal{E}} \psi^f_{xy}(S(x), S(y); \theta^f_2(I)) + \lambda_C \sum_{b=1}^{B} \psi_{C_b}(S(C_b); \theta_C(I, \mathbb{A})). \quad (9)$$

3. Experiments with brain MR data

To evaluate the performance and the robustness of the proposed label fusion method, we employ two available databases of T1-weighted (T1W) MR images: (i) 18 modified images from the Internet Brain Segmentation Repository (IBSR) [45, 46] and (ii) 50 images of epileptic and nonepileptic patients with hippocampal outlines (HFH) [47].

The IBSR contains images of healthy patients with expert segmentation of 43 anatomical structures. The voxel size of these images is 0.9375 x 1.5 x 0.9375 mm^3. In contrast, HFH contains a total of 50 images that were randomly divided into 25 images used for the training set and 25 used for the test set. Manual segmentations are only available for the training images. Images were acquired using two MR imaging systems with field strengths of 1.5 T and 3.0 T; thus, these images have different resolutions (0.78 x 2 x 0.78 mm^3 and 0.39 x 2 x 0.39 mm^3, respectively). Fig. 3 shows a coronal view of the T1W MR images together with the manual segmentations of similar brain locations for comparison. The intensity patterns and textures are quite different. The T1W MR images show large variations because not all of these images were acquired using the same scanner.

In the pre-processing of the databases, non-brain regions are removed from all structural images. Removing non-brain tissue prior to registration is generally accepted as a means to simplify the inter-subject registration

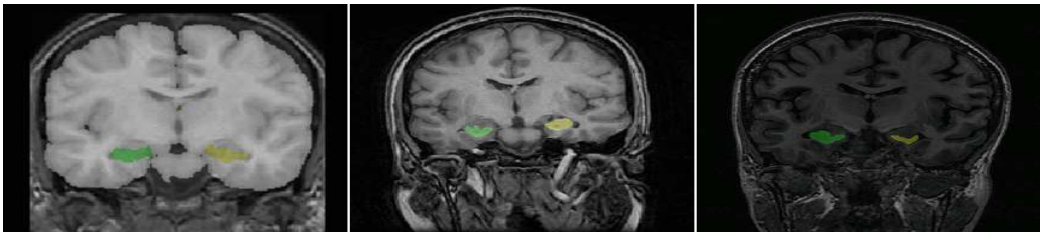

Figure 3: Coronal slices of similar brain locations for comparison: a) IBSR, b) HFH 1.5 T, and c) HFH 3.0 T

problem and thus increase the quality of the registrations [48, 49]. The images are skull-stripped using BET [50]. Then, all images are spatially normalized to a reference atlas using an affine registration. For the IBSR database, CMTK's affine registration tool is used [51]. In HFH, an atlas (HFH_021) is selected as a reference to which all atlases are then co-registered with an affine transformation using FLIRT with 12 degrees of freedom [52].

After spatial normalization for both of the databases, a region of interest is defined for each structure studied (left and right hippocampus) as the minimum bounding box containing the structure for all of the training atlases expanded by three voxels along each dimension. The patient image is also processed using a skull-striping filter and an affine transformation into the common reference space. Then, the normalized patient image is cropped around the structures of interest. For each ROI, the atlases are ranked based on their similarity according to the target image using the mutual information (MI) measure [53]. Then, the selected atlases are registered non-rigidly into the ROI of the target image. All non-rigid registrations are computed using *Elastix* [54], a publicly available package for medical image registration. The non-rigid registration of the images is based on the maximization of MI, in combination with a deformation field parameterized by cubic B-splines [55]. The MI is implemented according to [56], using a joint histogram size of 32 x 32 and cubic B-spline Parzen windows. A unique resolution is employed using a B-spline control point spacing of 3.0 mm in all directions. To optimize the cost function, an iterative stochastic gradient descent optimizer is used [57]. In each iteration, 2000 random samples are used to calculate the derivative of the cost function. A maximum of 500 iterations of the stochastic optimization procedure is used. The above-described settings were determined through trial-and-error experiments on two image pairs. These parameters of the non-rigid registrations are equally applicable to both databases.

The atlas-labeled images are modeled using the logarithm of odds (LogOdds) formulation, which is based on the signed distance transform [58]. This representation replaces the labels by the signed distances, which are assumed to be positive inside the structure of interest. We find that the LogOdds model produces more accurate results compared with trilinear interpolation or nearest-neighbor interpolation for transferring the atlas-labeled images [19].

Fig. 4 shows the relationship between the individual atlases and their performance in segmenting the target images. The DICE coefficient [59] is selected as a measure of the segmentation overlaps. The results are shown as the distributions of $DICE(\tilde{S}_i, S_R)$, where S_R is the ground-truth segmentation of the target image and i is the order of the atlas in the database from the similarity to the target image.

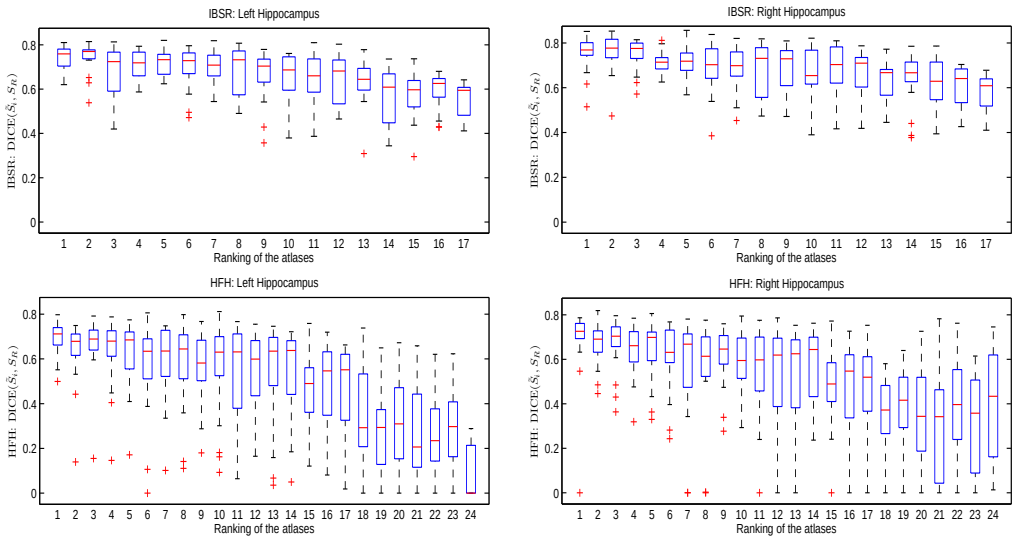

Figure 4: Relationship between the individual atlases and their performance in segmenting the target image. The graph illustrates the DICE distribution in segmenting by label propagation for a given rank (MI) and for an individual atlas, $DICE(\tilde{S}_i, S_R)$, where S_R is the ground-truth segmentation of the target image and i is the order of the atlas in the database from the similarity to the target image .

After the atlases are ranked and registered, we employ a leave-one-out validation strategy to determinate the number of atlases that are fused to

the target image [11]. The number of fused atlases depends on the label fusion method. Section 3.1.2 shows how to determine the number of atlases to fuse. Finally, the registered atlases are fused, the labeling is calculated based on graph cuts, and an inverse affine transformation is applied to return the segmentation into the native space of the target image. Fig. 5 shows a flow chart that summarizes the processing of the images.

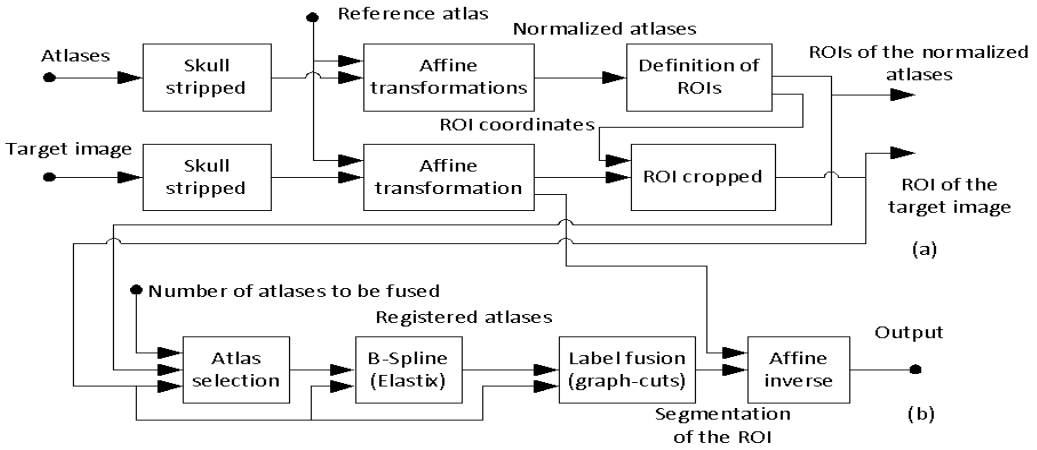

Figure 5: Flow chart summarizing the processing of the patient images: (a) Spatial normalization and definition of the ROIs. (b) Segmenting of the ROIs using image registrations and label fusion methods.

3.1. Setting CRF parameters

Given the CRF model in (9) and its parameters, the optimal labeling is found by applying the min-cut/max-flow algorithm of [60]. The parameters of the model for a ROI $\Theta = \{\theta_1, \theta_2, \theta_C\}$ are first learned by piecewise training and then recombined with the weighting multipliers $\Lambda = \{\lambda_R, \lambda_f, \lambda_C\}$. The CRF model is only trained for voxels whose labels have uncertainty such that the computational burden is reduced. A voxel is uncertain in its label when the atlas-labeled images are transferred and this voxel receives votes from different classes.

In the following subsections, we investigate the effects of different aspects of the model and then present the full quantitative and qualitative results for our approach and for other label fusion methods.

3.1.1. Learning the potentials

Image likelihood. We compare a generative appearance model and a discriminative appearance model. Using the generative model of (2), a pool of candidate features has to be defined. For each T1W MR image, the following features are calculated: intensity, gradients, Laplacians, curvatures and local entropies in different scales. Spatial derivatives are implemented by Gaussian-derivative filters. Some of these features are not invariant in gray level, and thus, an intensity normalization is applied to the registered atlas images using the histogram matching algorithm [61]. Given a set of extracted features from each voxel, a feature selection process is required. Because we use Bayesian classifiers and the result of the assignment is two labels, i.e., binary classification, an estimation of the Bayes error is given by the Bhattcharyya distance. Due to the ability to predict error using the Bhattcharyya distance, it is possible to determine the minimum number of features required for the classification tasks [62]. Then, each proposed subset of features by the Bhattcharyya distance is tested. The subset that maximizes the Dice coefficient using the generative classifiers will be used as the selected feature vector. These features are the intensity, the gradient norm with derivatives of Gaussians at scale 2 and the local entropy using a neighborhood kernel of size 9 x 3 x 9. To estimate the statistical parameters of the generative appearance model and because the number of samples for any label is low, a neighborhood system $\mathcal{N}(x)$ is tuned and applied to equations (3) and (4). $\mathcal{N}(x)$ is defined by a sphere with center x and radius of 1 mm. During runtime, a matrix of correlation coefficients is calculated in each voxel over the selected features. The scalar features, whose correlation coefficients are less than 0.6 in absolute value, are considered independents, and they are used in the unary potentials. Therefore, the dimension of the feature space is variable for each voxel and can be $d = 3, 2$ or 1.

Regarding the discriminative model, a 12-dimensional filter bank is applied to the registered atlases, generating a kd-tree model [44]. The registered atlas images are convolved with Gaussians at scales of 1, 2 and 4; derivatives of Gaussians at scales of 2 and 4; and Laplacians of Gaussians at scales of 1, 2 and 4 [37]. Given the 12-D responses of a voxel belonging to the target image, the training vectors in the kd-tree that are nearest are found. These vectors are used to calculate the distances to each label, and then equation (5) is applied to obtain $p(I(x)|l; \mathbb{A})$. The amount of training data in the discriminative model is often biased toward the background class. A classi-

fier learned using these data will have a prior preference for this class. To normalize for this bias, we weight each training example by the inverse class frequency. The classifiers trained using this weighting tend to provide better performance [63].

Label prior. In the weighted voting method for estimating the label prior probabilities, similarity measures are needed between regions of the target image and each registered atlas image, i.e., $m(I, \tilde{I}_i, x)$ in (6). Because a statistical relationship is assumed among the intensities of these images, MI is used as the similarity measure. The gain exponent is set to $q = 4$ [18]. A semi-global strategy is used to calculate the weight for each registered atlas. This strategy is most appropriate when the contrast between neighboring structures is low, as in the case of the hippocampus [18]. A binary mask is used to measure this similarity between the target image and the registered atlases. This mask is constructed by joining all transferred labeled images. A voxel is considered in the binary mask if at least a vote of the foreground class is received.

Spatial regularization. In the pairwise potentials and considering 3D grid-graphs with 6 neighborhood systems in $\mathcal{E}$, $\nabla I(x)$ is calculated by derivatives of Gaussians at scale 1 and applied to the Riemann potentials. The two types of the image-based Riemannian metrics have been implemented. The anisotropic metrics provided better results than the isotropic one, but the improvements were insignificant. Therefore, we selected the anisotropic metrics in the results to be shown.

The vector field, which is used to determine the orientation of the surface between labels, is the same used in the Riemannian metrics, i.e., $\vec{v}(x) = \nabla I(x)$. The flux requires determining whether the object is dark or bright according to the background. We estimate whether a voxel is dark by the sign of the Laplacian of Gaussian at scale 2, whereas unary potentials are used to infer whether this voxel belongs to the object. The flux term is implemented via edges to the terminals allowed given an arbitrary vector field, and this potential can be submodular [35].

Higher-order potentials. The higher-order potentials require a clique set of the target image. The cliques are also obtained from voxels whose labels have uncertainty. For this purpose, we use textons [36, 37]. An unsupervised clustering method is performed using the above 12-D responses from registered atlas image voxels. We employ the Euclidean distance k-means

clustering algorithm, which can be made faster through the use of an accelerated version with simple patches [64]. This algorithm returns the cluster centroid locations. Each centroid is assigned to a texton. Next, each voxel of the target image is labeled with a texton using the nearest cluster centroid. A clique is formed by spatially connecting voxels that have equal textons. We investigated changes in the number of textons. The number of cliques remained approximately invariant when the number of textons is low. A high number of textons increased the computational cost without resulting in significant improvements. Thus, we used 10 textons in our experiments. Fig 6 shows an example of cliques obtained using a ROI belonging to HFH with 10 textons.

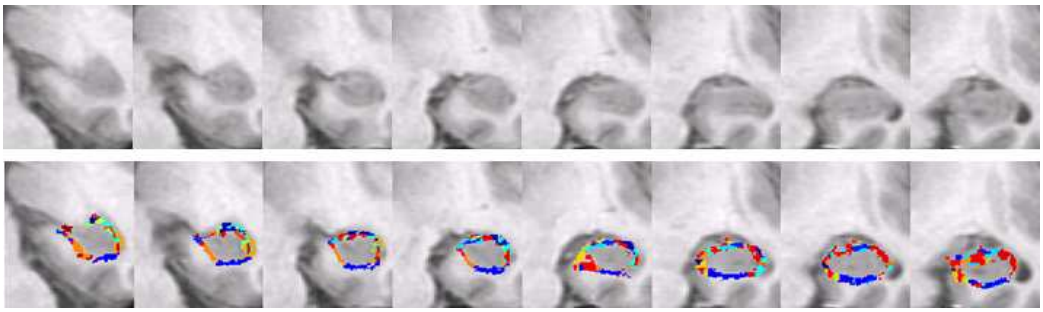

Figure 6: From a coronal view of a ROI belonging to HFH, the obtained cliques are shown using 10 textons. Each clique is labeled with a color. The upper row plots some slices of the left hippocampus. The lower row presents the slices with the overlapped cliques in colors.

The non-dominant label of each clique and its probability are inferred from (7). The weights ω_i^l are obtained using the minimum Euclidean distance between the i-voxel of the clique and voxels whose labels do not have uncertainty, e.g., ω_i^1 is the minimum distance from i-voxel to another that belongs to the k-structure without uncertainty. Then, these weights are normalized $\omega_i^l \leftarrow \frac{\omega_i^l \cdot \#C}{\sum_i \omega_i^l} G(C)$.

Finally, the four terms of the CRF model are combined by weighting multipliers $\Lambda = \{\lambda_R, \lambda_f, \lambda_C\}$, and they are tuned by Dice evaluation. These multipliers are varied in certain ranges, and their effects are measured from the overlap between the resulting segmentation and the ground truth. The multiplier vector is adjusted to provide the highest Dice coefficient values.

3.1.2. The number of fused atlases

The number of fused atlases depends on the label fusion method. We compare five label fusion methods: STAPLE, MV, WV and the two methods that we derive from our proposal (i.e., with the generative or discriminative appearance model). STAPLE estimates the performance of each transferred atlas-labeled image iteratively. STAPLE treats the label fusion as a maximum-likelihood problem and solves it using the expectation-maximization algorithm [17]. In MV, the transferred atlas-labeled images are equally weighted. For each voxel, the label with largest agreement from all registered atlases is assigned as the final label. A natural extension of MV is to improve from simple averaging to adaptive weighted averaging. In [18], various weighting strategies were categorized into two groups: (i) global weighted voting and (ii) local weighted voting. These authors also showed that the global weighted method outperforms the local solution when segmenting low-contrast brain structures. We apply to WV the same parameters with respect to our proposal in label prior (semi-global, MI and $q = 4$).

Once the parameters of the label fusion methods are defined, we employ a leave-one-out validation strategy to determinate the number of atlases that are fused to the target image for each label fusion method [11]. In our label fusion methods, we only use the unary potentials in the proposed CRF model. The unary potentials depend on the number of fused atlases. By contrast, the pairwise potentials depend exclusively on the target image, and the higher-order potentials depend on the quality of the unary potentials, the cliques and the target image. Using this procedure, we decouple the determination of the number of the atlases to be fused and the tuning of Λ in our proposal.

In all label fusion methods, the atlases are ranked using the MI measure between the target image and each atlas image. The i-first atlases are non-rigidly registered into the target image. Fig. 7 shows how the segmentation accuracy varies with the number of fused atlases. Each plotted point shows the average DICE coefficient in segmenting all of the target images for the number of fused atlases. We observe a difference between fusion methods that only use the transferred labeled atlases (STAPLE and MV) and those that employ all information of the registered atlases and the target image (WV and our unary potentials). As previously reported in [18], the STAPLE-based fusion rule does not necessary lead to higher Dice coefficients compared to the majority voting rule. Our unary potentials outperform WV, particularly with the discriminative appearance model. We observe that our

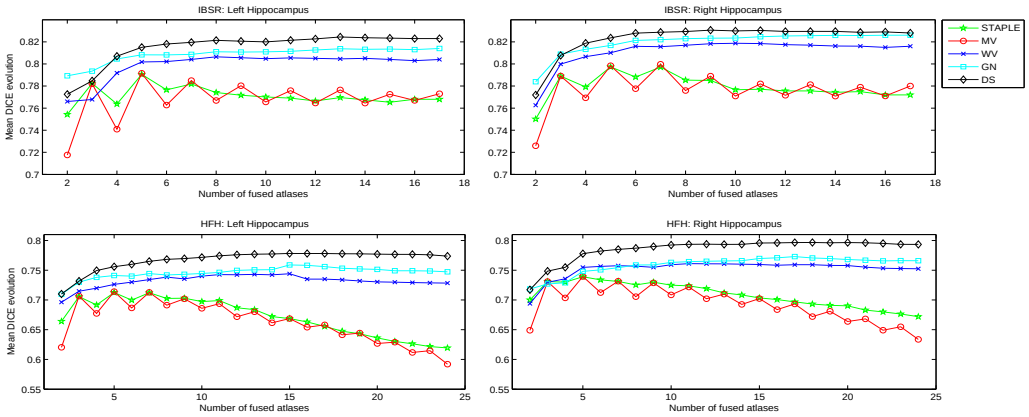

Figure 7: Relationship between the segmentation accuracy and the number of fused atlases. Plots show the mean DICE in all target images against the different number of fused atlases and for the five label fusion methods: STAPLE, MV, WV, unary potentials using the generative model (GN) and the discriminative model (DS).

appearance models are able to work with the label prior model and improve the segmenting results.

From this experiment, we fix the number of the fused atlases for each label fusion method as follows: a) STAPLE: 5, b) MV: 5, c) WV: 15, d) Generative CRF: 15, and e) Discriminative CRF: 15.

3.1.3. The effects of different potentials

We investigate the segmentation results on the IBSR and HFH databases when the potentials of the proposed CRF model are modified. We compare between the generative and discriminative appearance models combined with the other potentials.

Table 1 presents the quantitative segmentation results for the generative and discriminative models on the IBSR and HFH databases. Five combinations of the proposed CRF model are evaluated: a) only with unary potentials, b) unary and pairwise potentials without the influence of flux, c) unary and pairwise potentials with the influence of flux, d) unary and higher potentials and e) the full CRF model. The weighting multipliers are also reported for each case in the table. There are several conclusions. (1) The unary potentials are the most powerful in the classification task. (2) The pairwise and higher potentials improve the segmentations, making them

Table 1: Average values and standard deviations of the DICE coefficients for all training images belonging to IBSR/HFH data using the generative/discriminative appearance models with different combinations of potentials: a) $\lambda_R = \lambda_f = \lambda_C = 0$, b) $\lambda_f = \lambda_C = 0$, $\lambda_R = \{0.5, 0.25, 0.5, 0.5\}$, c) $\lambda_C = 0$, $\lambda_R = \{0.5, 0.25, 0.75, 0.5\}$, $\lambda_f = \{0.05, 0.025, 0.025, 0.025\}$ d) $\lambda_R = \lambda_f = 0$, $\lambda_C = \{0.2, 0.2, 0.75, 0.25\}$, e) $\lambda_R = \{0.5, 0.2, 0.75, 0.4\}$, $\lambda_f = \{0.05, 0.025, 0.025, 0.02\}$, $\lambda_C = \{0.5, 0.2, 0.25, 0.6\}$. The four elements of λ are the values that correspond to the use of the generative/discriminative appearance models for IBSR and HFH data, respectively.

Type		**IBSR**		**HFH**	
		Generative	Discriminative	Generative	Discriminative
a) Unary	LH	0.805 ± 0.042	0.822 ± 0.044	0.749 ± 0.076	0.778 ± 0.065
Potentials	RH	0.814 ± 0.055	0.828 ± 0.055	0.759 ± 0.060	0.795 ± 0.034
b) Unary +	LH	0.810 ± 0.042	0.825 ± 0.042	0.757 ± 0.078	0.780 ± 0.066
Riemann	RH	0.821 ± 0.049	0.830 ± 0.056	0.772 ± 0.060	0.797 ± 0.035
c) Unary +	LH	0.810 ± 0.042	0.826 ± 0.042	0.758 ± 0.078	0.779 ± 0.066
Pairwise	RH	0.821 ± 0.049	0.830 ± 0.055	0.774 ± 0.060	0.797 ± 0.034
d) Unary +	LH	0.809 ± 0.039	0.826 ± 0.043	0.751 ± 0.076	0.780 ± 0.066
H. potentials	RH	0.817 ± 0.052	0.831 ± 0.056	0.763 ± 0.058	0.797 ± 0.034
e) **Proposed**	LH	0.814 ± 0.039	0.826 ± 0.042	0.759 ± 0.076	0.781 ± 0.066
CRF model	RH	0.820 ± 0.050	0.831 ± 0.055	0.775 ± 0.060	0.798 ± 0.032

more accurate and robust. These potentials result in the object contour being more accurately delineated. (3) The pairwise potentials that take into account the orientations of the segmentation surfaces do not result in significant improvements. (4) The performance of the proposed higher-order potentials is comparable to that of the pairwise potentials. (5) The full CRF model provides slight improvements to any of the previous combinations of the proposed CRF model. (6) To further generalize the model, the weighted multipliers Λ are tuned to the same values for both the left and right hippocampus. Note that the values of Λ are similar between the two databases: (i) with the generative appearance model, the values are $\Lambda_{IBSR} = \{0.5, 0.05, 0.5\}$ and $\Lambda_{HFH} = \{0.75, 0.025, 0.25\}$), and (ii) with the discriminative appearance model, the values are $\Lambda_{IBSR} = \{0.2, 0.025, 0.2\}$ and $\Lambda_{HFH} = \{0.4, 0.02, 0.6\}$). The proposed CRF model is shown to be robust to variability in the databases.

Fig. 8 shows the hippocampal segmentation results for the best, one mean, and the worst subject using the proposed CRF with the discriminative appearance model.

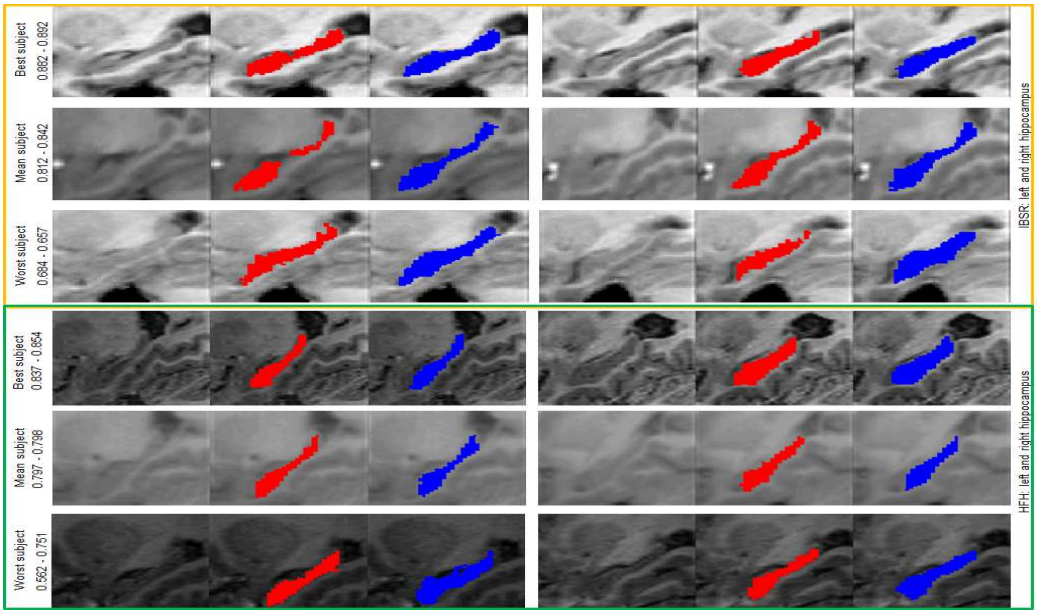

Figure 8: Left and right hippocampal segmentation for the subject with the best Dice coefficient (top), a mean Dice coefficient (middle), and the worst case (bottom) obtained using our proposed CRF model with the discriminative appearance model. The manual segmentations are shown in red, and the segmentations obtained with our method are shown in blue.

3.2. Results

Given the number of fused atlases and the parameters of the label fusion methods, Table 2 presents the quantitative segmentation results for each label fusion method and each ROI on the IBSR and HFH databases.

Statistical significance is evaluated using the Wilcoxon signed-rank test, where a $p-$value < 0.05 shows significant improvement. Given the DICE coefficient distributions of the full CRF model with the discriminative appearance model as references, the $p-$values are shown in Table 3 for the DICE coefficient distributions corresponding to the other label fusion methods. These values indicate significant improvement between our approach and other conventional approaches. Furthermore, note that there is no significant improvement when the discriminative model is replaced by the generative model.

Figure 9 presents the qualitative results of the five label fusion methods in terms of surface renderings of the left and right hippocampus for a subject. We have selected a representative patient for each database. Each

Table 2: Average values and standard deviations of the DICE coefficients for all training images belonging to IBSR/HFH data using the following approaches: STAPLE, MV, WV, full CRF with the generative appearance model and the discriminative model.

Type		**IBSR**	**HFH**
STAPLE	LH	0.793 ± 0.040	0.726 ± 0.120
	RH	0.804 ± 0.057	0.742 ± 0.063
Majority	LH	0.791 ± 0.040	0.714 ± 0.127
Voting	RH	0.798 ± 0.052	0.739 ± 0.074
Weighted	LH	0.805 ± 0.042	0.744 ± 0.067
Voting	RH	0.817 ± 0.053	0.760 ± 0.060
Generative	LH	0.814 ± 0.039	0.759 ± 0.076
CRF model	RH	0.820 ± 0.050	0.775 ± 0.060
Discriminative	LH	0.826 ± 0.042	0.781 ± 0.066
CRF model	RH	0.831 ± 0.055	0.798 ± 0.032

Table 3: $p-$values using the DICE coefficient distributions of the full CRF model with the discriminative appearance model as a reference.

Approach		**IBSR**	**HFH**
STAPLE	LH	$p = 0.006$	$p = 0.008$
	RH	$p = 0.08$	$p = 0.0002$
MV	LH	$p = 0.007$	$p = 0.0006$
	RH	$p = 0.02$	$p = 0.0004$
WV	LH	$p = 0.025$	$p = 0.01$
	RH	$p = 0.14$	$p = 0.008$
Generative	LH	$p = 0.12$	$p = 0.25$
	RH	$p = 0.15$	$p = 0.14$

segmentation strategy is compared to the manual gold standard labels. False positive voxels are drawn in green, and false negative voxels are shown in red. Visually, one can observe less red and green regions in the fusion results, indicating better segmentations. Our approach presents fewer false positives and negatives compared to other fusion methods. These images also indicate that the resolutions and manual protocols are different in each database [46, 47].

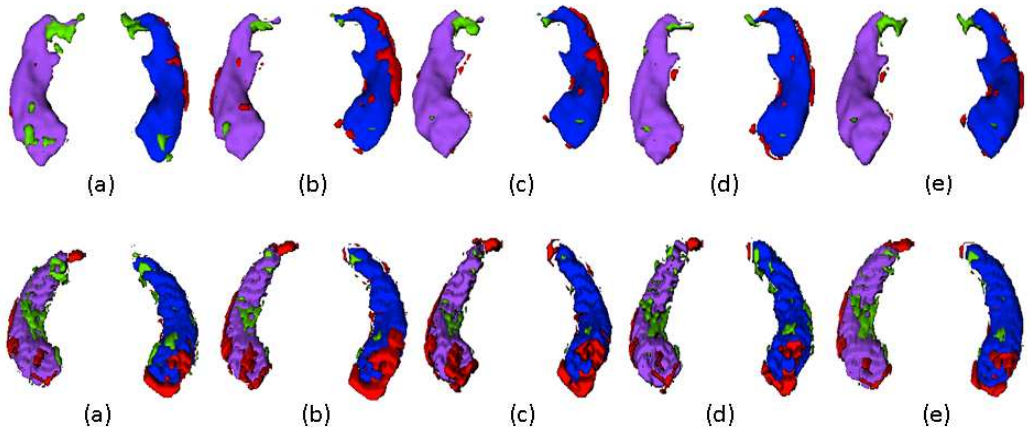

Figure 9: Results for the (a) STAPLE, (b) majority voting, (c) weighted voting, (d) generative approach and (e) discriminative approach. The images show surface renderings of the left and right hippocampus in the same patient for each database. The first row corresponds to a subject belonging to IBSR data, and the second row corresponds to HFH data. The manually labeled structures are rendered in purple and blue. Green regions indicate false positives, and red regions indicate false negatives.

The evaluation of the test images belonging to the HFH database is performed by an external team by submitting the results to a web site [47]. The given evaluations are of the entire hippocampus (see Table 4). These results are consistent with the values obtained in the training images of the HFH database.

Comparing segmentation results between different reported methods is always difficult. The quality of the databases used for validation, the anatomical definition of the structure, the quality of expert segmentations, the populations studied and the different measures reported all make comparison of the results difficult. With these caveats in mind, we compare our segmentation results with other approaches that used the same databases. Liu et al. [29] developed an auto context model to segment the sub-cortical structures from T1W MR images. Their technique combines a discriminative

Table 4: Average values and standard deviations of the DICE coefficients for all 25 test images belonging to HFH data using the proposed CRF model with the following appearance models: a) generative and b) discriminative.

Type		**Dice**
Generative	LH+RH	0.760 ± 0.053
Discriminative	LH+RH	0.778 ± 0.047

model for appearance with a label prior term. They tested their approach on IBSR data and compared their approach with FreeSurfer [42], which has been widely in this field. They reported average Dice coefficients of 0.75 and 0.74 for the hippocampus in FreeSurfer and their approach, respectively. The IBSR data were also used for the weighted voting method proposed by Artaechevarria et al. [18]. The best average Dice coefficients were 0.74 and 0.76 for the left and right hippocampus, respectively.

Jafari-Khouzani et al. [47] developed the HFH data. They evaluated two approaches on the HFH database: (i) Parser [8] and (ii) classifier fusion and labeling (CFL) [11]. Brain Parser uses Adaboost to select and fuse a set of features from the training data to obtain the discriminative appearance model. It is combined with a generative shape model. Jafari-Khouzani et al. reported an average Dice coefficient of 0.64 for the hippocampus. In CFL, the selected atlases are co-registered to the target image, and their transferred labels are fused using the vote rule. The authors reported a Dice coefficient of 0.75 for the hippocampus. Therefore, our results are as good as or even better than those previously reported. Table 5 shows the results reported in the literature obtained on the IBSR and HFH datasets. These comparatives show that the proposed approach is very competitive with respect to recently reported methods.

Table 6 presents the average values and standard deviations of the computing times in seconds for all training images. We only report the times for non-rigid registrations of the images and label fusions. The computational complexity is primarily due to the registration of the selected atlases into the target image. The computational time for segmentation increases linearly with the number of atlases that have to be registered. However, due to the availability and low cost of multi-core processors, this approach

Table 5: Comparison of the proposed method with other segmentation methods using the mean DICE coefficient on IBSR and HFH data.

	Proposed CRF model	Brain Parser [8, 47]	CFL [11, 47]
LH - RH	0.781-0.798	0.64	0.75

	Proposed CRF model	Fischl et al [42]	Liu et al [29]	Artaechavarria et al [18]
LH - RH	0.826-0.831	0.75	0.74	0.74-0.76

Table 6: Average values and standard deviations of the computing times in seconds for all training images ([Dual CPU] Intel Xeon E5520 @ 2.27 GHz)

Type		**IBSR**	**HFH**
Registration	LH	30.18 ± 1.74	41.63 ± 0.61
	RH	31.82 ± 1.99	39.28 ± 0.69
Label fusion	LH	9.46 ± 1.09	39.53 ± 6.93
	RH	9.21 ± 0.76	38.06 ± 5.98

is becoming more feasible. The task of non-rigid registrations has been parallelized. The registration of the 15 first atlases in a ROI requires less than 45 seconds. The computing times of the registration tasks have a weak dependence on the image resolution because B-Spline uses an isotropic grid with the same physical units (i.e., the spacing is specified in millimeters). By contrast, the computational cost of the label fusion method depends on the image resolution. Our approach takes an average of approximately 80 and 160 seconds for fusing labels (included non-rigid registrations) of both the left and right hippocampus on images belonging to IBSR and HFH, respectively. The code of the label fusion methods is not yet optimized; thus, the computing times can be easily reduced. The scripts used in this study are available at `https://www.nitrc.org/projects/lf_crf/`.

4. Discussion and conclusion

We introduce a label fusion method that is based on a CRF model and on ROIs. After dividing the image into anatomically meaningful regions, a CRF model is used to fuse the registered atlases in each ROI. The CRF model is characterized by a pseudo-Boolean function defined on unary, pairwise and higher-order potentials. The unary potentials combine an appearance model based on multiple features with a label prior using a weighted voting method. We compare a generative appearance model with a discriminative appearance model. The discriminative appearance model provides better results than the generative model, but it does not result in significant improvements. During the experiments, our appearance models are able to work with the label prior model and improve the segmentation results. Unary potentials were determined to be the most powerful term in the CRF model. An image-based Riemannian metric and the orientation of the segmentation surface are used to define pairwise potentials. The defined anisotropic Riemannian metric is selected to provide experimental results that are better than those provided by the isotropic metric. The orientation term of the segmentation surface is solved by estimating whether the object is darker or brighter than the background. The signs of the Laplacians of Gaussians and unary potentials are used to infer whether a voxel is dark or bright and whether it belongs to the object. Unary and pairwise potentials are coupled with higher-order potentials. The proposed higher-order potentials enforce label consistency in the cliques according to the estimations of unitary potentials and Euclidean distances. The cliques are clustered by spatially connecting voxels that have equal textons. After inferring the probabilities of each label in a clique, this potential family attempts to split the voxels of this clique between the labels related to these probabilities and the Euclidean distances between the voxels of this clique and the labeled regions without uncertainty.

These higher-order potentials only work if the cliques are homogeneous in appearance terms. Precisely, these cliques have greater difficulty in being correctly classified because the appearance models are not able to discern between labels. On the other hand, the cliques are only extracted in voxels where there are discrepancies among the transferred labeled atlases. These drawbacks are overcome by adding extra information through the higher-order potentials: (i) collective estimations that the cliques belong to a label using the unary potentials and (ii) the Euclidean distances between voxels of a clique and the labeled regions without uncertainty.

The proposed higher-order potentials are representable by graphs. An auxiliary binary variable is added to the graph for each clique. In our experiments, the performance of the new higher-order potentials is comparable to that of the pairwise potentials. Finally, a globally optimal binary labeling is found using a st-mincut algorithm in each ROI.

During our experiments, we applied the same parameters of the non-rigid registrations to both databases. To further generalize the proposed CRF model, the weighted multipliers Λ were tuned to the same values for both the left and right hippocampus. Note that the values of Λ are similar between the two databases. The proposed CRF model is shown to be robust to variability in the databases.

We compare our approach with other label fusion methods in the automatic hippocampal segmentation from T1W-MR images. We demonstrate that the proposed approach is very competitive with respect to recently reported methods.The scripts used in this study are available at `https://www.nitrc.org/projects/lf_crf/`.

The label fusion problem using the proposed CRF model reveals several points. (1) The k-NN appearance model could be replaced by other discriminative approaches, such as sophisticated randomized trees or boosting-based classifiers. Preliminary results on random forest [65] have shown similar results but with a higher computational cost (note that the discriminative model must also be trained at runtime). (2) With our experiments, we can not draw definitive conclusions regarding the proposed higher-order potentials. Although their performances are comparable to those of the pairwise potentials, more experiments are required for validation. This potential family may exhibit better performance in other contexts. (3) Although the label prior term imposes a prior shape, these constraints are reflected in a unary potential. The prior shape constraints should also be formulated in pairwise and higher-order potentials [66, 67]. We leave this issue for future work.

Another outstanding issue is the patch-based labeling approach. We believe that both approaches, label fusion methods using non-rigid registrations and patch-based labeling methods, are complementary and could work together to improve the results based on multi-atlas segmentation.

References

[1] Apostolova LG, Dutton RA, Dinov ID, Hayashi KM, Toga AW, Cummings JL, et al. Conversion of mild cognitive impairment to alzheimer

disease predicted by hippocampal atrophy maps. Archives of Neurology 2006;63(5):693–9.

[2] Berretta S, Pantazopoulos H, Lange N. Neuron numbers and volume of the amygdala in subjects diagnosed with bipolar disorder or schizophrenia. Biological Psychiatry 2007;62(8):884–93.

[3] Geuze E, Vermetten E, Bremner J. MR-based in vivo hippocampal volumetrics: 1. Review of methodologies currently employed. Molecular Psychiatry 2004;10(2):147–59.

[4] Nestor SM, Gibson E, Gao FQ, Kiss A, Black SE. A direct morphometric comparison of five labeling protocols for multi-atlas driven automatic segmentation of the hippocampus in Alzheimer's disease. NeuroImage 2013;66:50–70.

[5] Apostolova LG, Thompson PM. Brain mapping as a tool to study neurodegeneration. Neurotherapeutics 2007;4(3):387–400.

[6] Chupin M, Mukuna-Bantumbakulu AR, Hasboun D, Bardinet E, Baillet S, Kinkingnéhun S, et al. Anatomically constrained region deformation for the automated segmentation of the hippocampus and the amygdala: Method and validation on controls and patients with Alzheimer's disease. Neuroimage 2007;34(3):996–1019.

[7] Cremers D, Rousson M. Efficient kernel density estimation of shape and intensity priors for level set segmentation. In: Suri JS, Farag A, editors. Parametric and Geometric Deformable Models: An application in Biomaterials and Medical Imagery. Springer; 2007, p. 447–60.

[8] Tu Z, Narr K, Dollár P, Dinov I, Thompson P, Toga A. Brain anatomical structure segmentation by hybrid discriminative/generative models. IEEE Transactions on Medical Imaging 2008;27(4):495–508.

[9] Patenaude B, Smith SM, Kennedy DN, Jenkinson M. A bayesian model of shape and appearance for subcortical brain segmentation. Neuroimage 2011;56(3):907–22.

[10] Babalola KO, Patenaude B, Aljabar P, Schnabel J, Kennedy D, Crum W, et al. An evaluation of four automatic methods of segmenting the subcortical structures in the brain. Neuroimage 2009;47(4):1435–47.

[11] Aljabar P, Heckemann R, Hammers A, Hajnal J, Rueckert D. Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. Neuroimage 2009;46(3):726–38.

[12] Barnes J, Foster J, Boyes R, Pepple T, Moore E, Schott J, et al. A comparison of methods for the automated calculation of volumes and atrophy rates in the hippocampus. Neuroimage 2008;40(4):1655–71.

[13] Rohlfing T, Brandt R, Menzel R, Russakoff D, Maurer C. Quo vadis, atlas-based segmentation? Handbook of Biomedical Image Analysis 2005;:435–86.

[14] 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):115–26.

[15] Lotjonen J, Wolz R, Koikkalainen J, Thurfjell L, Waldemar G, Soininen H, et al. Fast and robust multi-atlas segmentation of brain magnetic resonance images. Neuroimage 2010;49(3):2352–65.

[16] Collins DL, Pruessner JC. Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion. Neuroimage 2010;52(4):1355–66.

[17] Warfield S, Zou K, Wells W. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging 2004;23(7):903–21.

[18] Artaechevarria X, Muñoz-Barrutia A, Ortiz-de Solorzano C. Combination strategies in multi-atlas image segmentation: Application to brain MR data. IEEE Transactions on Medical Imaging 2009;28(8):1266–77.

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

[20] Coupé P, Manjón JV, Fonov V, Pruessner J, Robles M, Collins DL. Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. Neuroimage 2011;54(2):940–54.

[21] Rousseau F, Habas PA, Studholme C. A supervised patch-based approach for human brain labeling. IEEE Transactions on Medical Imaging 2011;30(10):1852–62.

[22] Wang H, Suh JW, Das SR, Pluta JB, Craige C, Yushkevich PA. Multi-atlas segmentation with joint label fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence 2013;35(3):611–23.

[23] Tong T, Wolz R, Coupé P, Hajnal JV, Rueckert D. Segmentation of mr images via discriminative dictionary learning and sparse coding: application to hippocampus labeling. Neuroimage 2013;76:11–23.

[24] Wu G, Wang Q, Zhang D, Nie F, Huang H, Shen D. A generative probability model of joint label fusion for multi-atlas based brain segmentation. Medical Image Analysis 2014;18(6):881–90.

[25] van der Lijn F, den Heijer T, Breteler M, Niessen W. Hippocampus segmentation in MR images using atlas registration, voxel classification, and graph cuts. Neuroimage 2008;43(4):708–20.

[26] Buades A, Coll B, Morel J. A review of image denoising algorithms, with a new one. Multiscale Modeling and Simulation 2006;4(2):490–530.

[27] Han X, Fischl B. Atlas renormalization for improved brain MR image segmentation across scanner platforms. IEEE Transactions on Medical Imaging 2007;26(4):479–86.

[28] Shi F, Yap PT, Fan Y, Gilmore JH, Lin W, Shen D. Construction of multi-region-multi-reference atlases for neonatal brain MRI segmentation. Neuroimage 2010;51(2):684–93.

[29] Liu CY, Iglesias JE, Toga A, Tu Z. Fusing adaptive atlas and informative features for robust 3D brain image segmentation. In: IEEE International Symposium on Biomedical Imaging: From Nano to Macro. 2010, p. 848–51.

[30] Morra JH, Tu Z, Apostolova LG, Green AE, Toga AW, Thompson PM. Comparison of adaboost and support vector machines for detecting alzheimer's disease through automated hippocampal segmentation. IEEE Transactions on Medical Imaging 2010;29(1):30–43.

[31] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: International Conference on Machine Learning. 2001, p. 282–9.

[32] Tu Z. Auto-context and its application to high-level vision tasks. In: IEEE Conference on Computer Vision and Pattern Recognition. 2008, p. 1–8.

[33] Li Q, Wang J, Tu Z, Wipf DP. Fixed-point model for structured labeling. In: Proceedings of the 30th International Conference on Machine Learning (ICML-13). 2013, p. 214–21.

[34] Kohli P, Torr PH. Robust higher order potentials for enforcing label consistency. International Journal of Computer Vision 2009;82(3):302–24.

[35] Kolmogorov V, Boykov Y. What metrics can be approximated by geo-cuts, or global optimization of length/area and flux. In: Tenth IEEE International Conference on Computer Vision; vol. 1. 2005, p. 564–71.

[36] Leung T, Malik J. Representing and recognizing the visual appearance of materials using three-dimensional textons. International Journal of Computer Vision 2001;43(1):29–44.

[37] Shotton J, Winn J, Rother C, Criminisi A. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. International Journal of Computer Vision 2009;81(1):2–23.

[38] Song Z, Tustison N, Avants B, Gee J. Integrated graph cuts for brain MRI segmentation. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006;4191:831–8.

[39] Wolz R, Heckemann RA, Aljabar P, Hajnal JV, Hammers A, Lötjönen J, et al. Measurement of hippocampal atrophy using 4D graph-cut segmentation: application to ADNI. Neuroimage 2010;52(1):109–18.

[40] Ashburner J, Friston K. Unified segmentation. Neuroimage 2005;26(3):839–51.

[41] Pohl K, Fisher J, Grimson W, Kikinis R, Wells W. A bayesian model for joint segmentation and registration. Neuroimage 2006;31(1):228–39.

[42] Fischl B, Salat D, Busa E, Albert M, Dieterich M, Haselgrove C, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 2002;33(3):341–55.

[43] Raudys S, Jain A. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. IEEE Transactions on Pattern Analysis and Machine Intelligence 1991;13(3):252–64.

[44] Mount DM, Arya S. Ann: A library for approximate nearest neighbor searching. `http://www.cs.umd.edu/~mount/ANN/`; 2010 (Accessed: 20 January 2015). Version 1.1.2.

[45] Internet brain segmentation repository, IBSR. `http://www.cma.mgh.harvard.edu/ibsr/`; (Accessed: 20 January 2015).

[46] Rohlfing T. Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. IEEE Transactions on Medical Imaging 2012;31(2):153–63.

[47] Jafari-Khouzani K, Elisevich KV, Patel S, Soltanian-Zadeh H. Dataset of magnetic resonance images of nonepileptic subjects and temporal lobe epilepsy patients for validation of hippocampal segmentation techniques. Neuroinformatics 2011;9(4):335–46.

[48] Battaglini M, Smith SM, Brogi S, De Stefano N. Enhanced brain extraction improves the accuracy of brain atrophy estimation. Neuroimage 2008;40(2):583–9.

[49] Klein A, Andersson J, Ardekani BA, Ashburner J, Avants B, Chiang MC, et al. Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration. Neuroimage 2009;46(3):786–802.

[50] Smith SM. Fast robust automated brain extraction. Human brain mapping 2002;17(3):143–55.

[51] Studholme C, Hill D, Hawkes D. An overlap invariant entropy measure of 3D medical image alignment. Pattern Recognition 1999;32(1):71–86.

[52] Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 2002;17(2):825–41.

[53] Viola P, Wells III WM. Alignment by maximization of mutual information. International Journal of Computer Vision 1997;24(2):137–54.

[54] Klein S, Staring M, Murphy K, Viergever M, Pluim J. Elastix: a toolbox for intensity-based medical image registration. IEEE transactions on Medical imaging 2010;29(1).

[55] Rueckert D, Sonoda L, Hayes C, Hill D, Leach M, Hawkes D. Nonrigid registration using free-form deformations: application to breast MR images. IEEE Transactions on Medical Imaging 1999;18(8):712–21.

[56] Thévenaz P, Unser M. Optimization of mutual information for multiresolution image registration. IEEE Transactions on Image Processing 2000;9(12):2083–99.

[57] Klein S, Staring M, Pluim J. Evaluation of optimization methods for nonrigid medical image registration using mutual information and b-splines. IEEE Transactions on Image Processing 2007;16(12):2879–90.

[58] Pohl KM, Fisher J, Shenton M, McCarley RW, Grimson WEL, Kikinis R, et al. Logarithm odds maps for shape representation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI. Springer; 2006, p. 955–63.

[59] Dice L. Measures of the amount of ecologic association between species. Ecology 1945;26(3):297–302.

[60] Boykov Y, Kolmogorov V. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Transactions on Pattern Analysis and Machine Intelligence 2004;26(9):1124–37.

[61] Gonzales RC, Woods RE. Digital image processing. New Jersey: Prentice Hall 2002;6:1–689.

[62] Choi E, Lee C. Feature extraction based on the Bhattacharyya distance. Pattern Recognition 2003;36(8):1703–9.

[63] Shotton J, Johnson M, Cipolla R. Semantic texton forests for image categorization and segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. 2008, p. 1–8.

[64] Bishop CM, Nasrabadi NM. Pattern Recognition and Machine Learning; vol. 1. Springer New York; 2006.

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

[66] Krähenbühl P, Koltun V. Efficient inference in fully connected CRFs with Gaussian edge potentials. Preprint arXiv:12105644 2012;.

[67] Rother C, Kohli P, Feng W, Jia J. Minimizing sparse higher order energy functions of discrete variables. In: IEEE Conference on Computer Vision and Pattern Recognition. 2009, p. 1382–9.