A GENERATIVE MODEL FOR MULTI-ATLAS SEGMENTATION ACROSS MODALITIES

Juan Eugenio Iglesias[1], *Mert Rory Sabuncu*[1] *and Koen Van Leemput*[1,2,3]

[1] Martinos Center for Biomedical Imaging, MGH, Harvard Medical School, USA
[2]Department of Informatics and Mathematical Modeling, Technical University of Denmark
[3]Departments of Information and Computer Science and of Biomedical
Engineering and Computational Science, Aalto University, Finland

ABSTRACT

Current label fusion methods enhance multi-atlas segmentation by locally weighting the contribution of the atlases according to their similarity to the target volume after registration. However, these methods cannot handle voxel intensity inconsistencies between the atlases and the target image, which limits their application across modalities or even across MRI datasets due to differences in image contrast. Here we present a generative model for multi-atlas image segmentation, which does not rely on the intensity of the training images. Instead, we exploit the consistency of voxel intensities within regions in the target volume and their relation to the propagated labels. This is formulated in a probabilistic framework, where the most likely segmentation is obtained with variational expectation maximization (EM). The approach is demonstrated in an experiment where T_1-weighted MRI atlases are used to segment proton-density (PD) weighted brain MRI scans, a scenario in which traditional weighting schemes cannot be used. Our method significantly improves the results provided by majority voting and STAPLE.

Index Terms— Label fusion, multi-atlas segmentation

1. INTRODUCTION

In traditional atlas-based segmentation[1, 2], a template volume (henceforth "atlas") is registered (i.e. spatially aligned) to a target scan. The resulting deformation is then used to propagate the associated manual labels (either deterministic or probabilistic) and obtain the final segmentation. The main disadvantage of this technique is that a single deformation is limited in representing the whole population of potential target cases. This issue can be addressed by registering multiple atlases to the target volume. The question is then how to combine the "opinions" of different atlases. This problem is known as label fusion.

Majority voting[3], in which the mode of the propagated labels is selected as the final segmentation for each voxel, is widely used in medical imaging because it is straightforward to implement and, thanks to the maturity of registration methods, often yields good results. However, majority voting might not be able to correctly segment features that are present but underrepresented in the training dataset. STAPLE[4] weighs the propagated labels according to an estimated accuracy level, while incorporating consistency constraints. Neither majority voting nor STAPLE consider image intensities after registration. Higher segmentation accuracy can be achieved by considering the intensity of the images in label fusion, giving higher weights to the atlases which are more similar to the target volume, locally or globally[5, 6].

Weighted label fusion relies on the consistency of voxel intensities across scans. This represents a limitation in MRI, in which intensities depend heavily on the pulse sequence, acquisition settings and hardware. Mutual information has successfully been used for registration in such scenarios, but its inherent non-locality limits its application for local label fusion. Intensity normalization can ameliorate this problem, but only if the atlases and the target image are not acquired with different MRI contrast (e.g. T_1- vs. T_2-weighted).

Here we present a generative model for multi-atlas segmentation, which exploits the intensities of the target volumes by modeling their consistency within the regions to be segmented and their relation to the propagated labels. The intensities of the deformed templates are not considered in label fusion, so the set of atlases can include different modalities. As in [6], the registrations are assumed to be known (i.e. regarded as preprocessing), and we define a latent field that assigns each target voxel to a training volume. We assume that the image intensities for each label are samples from a Gaussian mixture model (GMM), modulated by a low-frequency bias field. Segmentation is formulated as finding the most likely labels in this framework. The algorithm is related to classical single atlas models[7], with two major differences: 1. we use a subject-specific prior probability for the labels computed in target space (as opposed to atlas space); and 2. the probabilistic model for the labels has a different structure.

Research supported by the NIH NCRR (P41-RR14075), the NIBIB (R01EB006758, R01EB013565), the NINDS (R01NS052585), the Academy of Finland (133611), and the Finnish Funding Agency for Technology and Innovation. M.R. Sabuncu is also supported by a KL2 MeRIT grant (Harvard Catalyst, NIH #1KL2RR025757-01 and financial contributions from Harvard and its affiliations), and a NIH K25 grant (NIBIB 1K25EB013649-01).

Table 1. Equations corresponding to the model in Figure 1.

1. $M \sim \frac{1}{Z(\beta)} \prod_{x \in \Omega} \exp\left(\beta \sum_{y \in \mathcal{N}_x} \delta(M(x) = M(y))\right)$
2. $L(x) \sim \frac{\exp\left(\rho D_{M(x)}^{L(x)}(x)\right)}{\sum_{l'=1}^{\mathcal{L}} \exp\left(\rho D_{M(x)}^{l'}(x)\right)}$
3. $I^*(x) \sim \sum_{k=1}^{\mathcal{C}_{L(x)}} \frac{w_{L(x),k}}{\sqrt{(2\pi\sigma_{L(x),k}^2)}} \exp\left[-\frac{(I^*(x)-\mu_{L(x),k})^2}{2\sigma_{L(x),k}^2}\right]$
4. $I(x) = I^*(x) \exp\left[-\sum_p c_p \psi_p(x)\right]$

2. METHODS

2.1. Generative Model

The generative model is depicted in Figure 1, and the corresponding equations are listed in Table 1. $\{L_n\}$, $n = 1, \dots, N$, represents the propagated labels from N available registered atlases. For a given voxel $x \in \Omega$ (where Ω is the image domain), $L(x)$ is the underlying segmentation label we are trying to estimate, which takes discrete values between 1 and $\mathcal{L}$. $L(x)$ is generated by the $M(x)^{th}$ atlas through a logOdds model[8] (Equation 2 in Table 1). The slope ρ of the model is assumed to be fixed and known. $D_n^l(x)$ represents the signed distance transform for atlas n and label l.

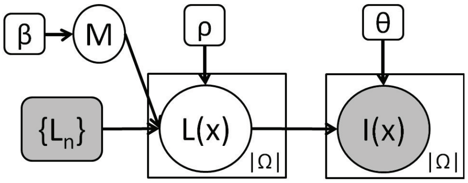

Fig. 1. Proposed graphical model. Plates indicate replication, shaded variables are observed.

The indices $M(x)$ are not independent across voxels, but distributed according to a Markov Random Field (MRF) with smoothness parameter β (see Equation 1 in Table 1, where δ represents Kronecker's delta and $\mathcal{N}_x$ is the 6-neighborhood of voxel x). Now, given the labels $L(x)$, the "true", underlying intensities $I^*(x)$ are assumed to be samples of a GMM with $\mathcal{C}_{L(x)}$ components and parameters $\Gamma_{L(x)} = \{\{w_{L(x),k}\}, \{\mu_{L(x),k}\}, \{\sigma^2_{L(x),k}\}\}$ (Equation 3 in Table 1). These underlying intensities are corrupted by a multiplicative bias field to yield the observed intensities $I(x)$ (Equation 4 in the table, where $\{\psi_p(x)\}$ is a set of spatially smooth basis functions). We use Θ to summarize all these image intensity parameters i.e., $\Theta = \{\{\Gamma_l\}, \{c_p\}\}$. The bias field basis functions $\{\psi_p(x)\}$, as well as the parameters β and $\{\mathcal{C}_l\}$ are assumed to be fixed and known. Note that the bias field is explicitly modeled as an exponential, which guarantees that it is greater than zero.

Some popular segmentation frameworks are particular cases of the proposed method. If $\beta = 0$, the model becomes very similar to a classical atlas-based segmentation method, e.g. [7], with the difference that here the label priors are built in subject space. Also, if $\beta = \infty$, the labels are generated only by the atlas that is closest to the target volume, which is close to best template selection. Another particular case of interest would be $\beta = 0, \sigma^2_{l,k} = \infty, \forall l, k$, in which case the algorithm simplifies to majority voting.

2.2. Segmentation

We formulate the segmentation as an optimization problem:

$$\begin{aligned}\{\hat{L}, \hat{\Theta}\} &= \arg\max_{L,\Theta} p(L, \Theta | I, \{L_n\}) \\ &= \arg\max_{L,\Theta} \log\left[p(I|L,\Theta)p(L|\{L_n\})\right]\end{aligned}$$

where we assume $p(\Theta) \propto 1$. We solve this problem using coordinate ascent, iteratively optimizing for L and Θ until convergence (often after 5-10 iterations).

2.2.1. *Optimizing Θ for fixed L*

The problem simplifies to $\hat{\Theta} = \arg\max_\Theta \log p(I|L,\Theta)$. Again, we use coordinate ascent within this step, alternatively optimizing for the parameters of the GMMs $\{\Gamma_l\}$ and the bias field $\{c_p\}$. Updating $\{\Gamma_l\}$ is closed-form and amounts to computing sample means and variances (if $\mathcal{C}_l = 1$) or a simple 1-D EM algorithm if $\mathcal{C}_l > 1$ (we use $\mathcal{C}_l = 1$ for all labels except the background). To update $\{c_p\}$, setting the derivatives equal to zero leads to a nonlinear system of equations that has no closed-form solution. Instead we use a numerical gradient ascent scheme with line search.

2.2.2. *Optimizing L for fixed Θ*

Here the problem is:

$$\begin{aligned}\hat{L} &= \arg\max_L \log\left[p(I|L,\Theta)p(L|\{L_n\})\right] \\ &= \arg\max_L \log\left[p(I|L,\Theta)\sum_M p(L|M,\{L_n\})p(M)\right]\end{aligned}$$

which is intractable because $M(x)$ includes interdependencies between neighboring voxels. Instead, we use variational EM to maximize a lower bound. Given that the Kullback-Leibler divergence $KL(A,B) \geq 0$, we define:

$$\begin{aligned}J &= \log\left[p(I|L,\Theta)p(L|\{L_n\})\right] - KL(q(M), p(M|I,\Theta,\{L_n\},L)) \\ &= H(q) + \sum_M q(M) \log p(M, L, I|\Theta, \{L_n\})\end{aligned}$$

where $H(A)$ represents the entropy and $q(M)$ is an arbitrary distribution on M. To maximize J, we use the standard computational trick that q is constrained to have the structure: $q(M) = \prod_{x\in\Omega} q_x(M(x))$. Then, variational EM maximizes J by iterating between an expectation (E) step which updates $q(M)$ and a maximization (M) step which optimizes L.

E step: in the E step, we look for the $q(M)$ which maximizes J, which happens when the KL divergence between $q(M)$ and the posterior $p(M|I, \Theta, \{L_n\}, L)$ is minimized:

$$\hat{q} = \arg\min_q \sum_M q(M) \log \frac{q(M)}{p(I, L|M, \Theta, \{L_n\})p(M)}$$

where $p(L, I|\Theta, \{L_n\})$ is omitted because it does not depend on M or q. Now, given the structure of $q(M)$ and also:

$$p(I, L|M, \Theta, \{L_n\}) = \prod_{x \in \Omega} p(I(x)|\Gamma_{L(x)}, \{c_p\})p(L(x)|L_{M(x)}(x))$$

it can be shown that the problem becomes:

$$\arg\min_q \mathbb{E}_q[-\log p(M)] + \sum_{x \in \Omega} \mathbb{E}_{q_x}\left[\log q_x(M(x))\right] - \dots$$

$$\dots - \mathbb{E}_{q_x}\left[\log p(I(x)|\Gamma_{L(x)}, \{c_p\}) + \log p(L(x)|L_{M(x)}(x))\right]$$

with the constraint that $q_x(m)$ lies on the probability simplex. Writing the Lagrangian, taking derivatives with respect to q_x and setting them to zero yields:

$$q_x(M(x)) = e^{\lambda(x)} p(L(x)|L_{M(x)}(x)) e^{\sum_{y \in \mathcal{N}_x} q_y(M(x))}$$

where $\lambda(x)$ is the multiplier that ensures $\sum_m q_x(m) = 1$. The equation can be solved with fixed point iterations.

M step: the goal of the M step is to maximize J with respect to L assuming $q(M)$ constant. Since we can discard the term $H(q)$, and because of the structure of $q(M)$, the maximization can be carried out voxel by voxel using exhaustive search over labels:

$$\hat{L}(x) = \arg\max_l \sum_{m=1}^{N} q_x(m) \log\left[p(I(x)|\Gamma_l, \{c_p\})p(l|L_m(x))\right]$$

As in most label fusion methods, the resulting label map is not guaranteed to preserve the topology of the different structures. However, this seldom happens in practice and can always be corrected with post-processing.

3. EXPERIMENTS AND RESULTS

3.1. Data

Two datasets (training and test) were used in this study. The test dataset consists of FLASH PD-weighted brain scans from eight healthy subjects (1.5T, TR=20ms, TE=min, $\alpha = 3°$, 1 mm. isotropic voxels). A total of 36 structures were manually labeled using the protocol described in [9]. The human raters took advantage of higher contrast, co-registered T_1-weighted data that we did not use in this study. As in [6], we only used a representative subset of the 36 structures for evaluation in this study: white matter (WM), cerebral cortex (CT), lateral ventricle (LV), hippocampus (HP), thalamus (TH), caudate (CA), putamen (PU), pallidum (PA), and amygdala (AM).

The training dataset consists of 39 T_1-weighted brain MRI scans (MP-RAGE, 1.5T, TR=9.7ms, TE=4.ms, TI=20ms, $\alpha = 10°$, 1 mm. isotropic resolution) and corresponding manual delineations of the same brain structures (same labeling protocol). We note that these are the same subjects that were used to construct the probabilistic segmentation atlas in FreeSurfer[10]. These scans were bias-field corrected and skull-stripped using FreeSurfer, and then ITK (`www.itk.org`) was used to register them to the PD-weighted data with a nonrigid transform (a grid of control points and b-splines) and mutual information as metric.

Table 2. Average Dice score across all structures for each method. The p-values correspond to a paired t-test which assesses whether the proposed method (at $\beta = 0.75$) produces a higher Dice score than each of the other algorithms.

Method	Maj.Vot.	STAPLE	$\beta = 0$	$\beta = 0.75$
Dice	0.766	0.762	0.796	0.802
p	2e-13	9e-22	3e-8	N/A

3.2. Experimental setup

We segmented the eight target volumes using majority voting, STAPLE, and the proposed method with $\beta = 0$ (local fusion) and $\beta = 0.75$, which represents the general case. We arbitrarily chose this value for β to match that from *Sabuncu et al*[6]. The same motivation was used to set $\rho = 1$. Three mixture components were used for the background class, and a single Gaussian for all other classes. The basis functions $\{\psi_p\}$ were set to a third order polynomial, which in 3D yields 20 coefficients. We set the number of atlases to $N = 15$; the presented results are an average over five runs with different, randomly selected atlases from the pool of 38. The accuracy of the segmentation is evaluated using Dice scores ($Dice(A, B) = \frac{2|A \cap B|}{|A| + |B|}$), and statistical significance is assessed with one-tailed, paired t-tests.

3.3. Results

Figure 2 shows the Dice scores for the different methods and structures of interest, whereas Table 2 shows the averages across all structures. STAPLE and majority voting provide comparable results. The proposed method (for $\beta = 0.75$) significantly outperforms STAPLE for all brain structures except for the putamen. Compared with majority voting, the scores when $\beta = 0.75$ are significantly higher for WM, CT, LV, HP and CA, though not significantly different for TH, PU, PA and AM. The hippocampus, pallidum and especially the amygdala seem particularly difficult to segment in the PD-weighted dataset, since there is very little image contrast around them. When we compare the results from our method for $\beta = 0$ and $\beta = 0.75$, the difference is not large in absolute terms (approximately 0.6%), but it is statistically very significant; the higher β seems to fix some consistent mistakes that are made when the voxels are considered independently. Finally, Figure 3 shows a sample coronal slice of a target scan, illustrating the segmentations given by majority voting and the proposed method, as well as the estimated bias field.

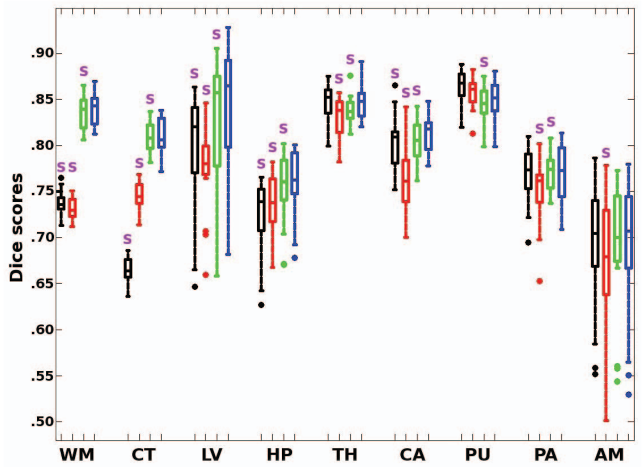

Fig. 2. Box plot of Dice scores for majority voting (black), STAPLE (red), $\beta = 0$ (green) and $\beta = 0.75$ (blue). The box has lines at the three quartile values. Whiskers extend to the most extreme values within 1.5 times the interquartile range from the ends of the box. Samples beyond those points are marked with circles. The purple S denotes that the scores are significantly lower than for our method at p=0.025.

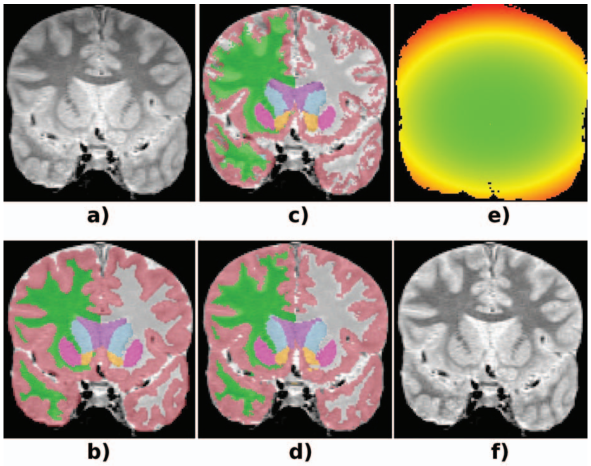

Fig. 3. Coronal slice of a test subject. a) Original. b) Ground truth segmentation. c) Majority voting. d) Proposed method ($\beta = 0.75$). e) Bias field estimated by the proposed method (green=1.0, red=1.25). f) Bias-field corrected.

4. CONCLUSION

A method to perform inter-modality label fusion has been presented in this paper. The method exploits the intensities of the target image and, despite the low number of test subjects, achieves statistical significance when compared with majority voting and STAPLE in a challeging PD-weighted dataset. The MRF prior has a positive effect on the peformance of the method, significantly improving the Dice scores. We hypothesize that, as the number of available atlases grows, increasing β would be beneficial. As $N \rightarrow \infty$, there would always be a training volume (almost) identical to the target case, and $\beta = \infty$ would provide a perfect segmentation.

The proposed inference method for the model can be improved by marginalizing over the labels when optimizing for the parameters, at the cost of introducing another level of complexity in the model. This will be addressed in future work, which will also include incorporating the registration in the optimization process, sweeping the parameters to assess their impact on segmentation accuracy, explicitly optimizing for ρ and testing the model on T_1 and multi-spectral data.

5. REFERENCES

[1] DL Collins, CJ Holmes, TM Peters, and AC Evans, "Automatic 3-d model-based neuroanatomical segmentation," *Hum. Brain. Mapp.*, vol. 3, pp. 190–208, 1995.

[2] BTT Yeo, MR Sabuncu, R Desikan, B Fischl, and P Golland, "Effects of registration regularization and atlas sharpness on segmentation accuracy," *Med. Image. Anal.*, vol. 12, no. 5, pp. 603–615, 2008.

[3] RA Heckemann, JV Hajnal, P Aljabar, D Rueckert, and A Hammers, "Automatic anatomical brain mri segmentation combining label propagation and decision fusion," *NeuroImage*, vol. 33, no. 1, pp. 115–126, 2006.

[4] SK Warfield, KH Zou, and WM Wells, "Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation," *IEEE Trans. Med. Im.*, vol. 23, no. 7, pp. 903–921, 2004.

[5] I Isgum, M Staring, A Rutten, M Prokop, MA Viergever, and B van Ginneken, "Multi-atlas-based segmentation with local decision fusion: application to cardiac and aortic segmentation in ct scans," *IEEE Trans. Med. Im.*, vol. 28, no. 7, pp. 1000–1010, 2009.

[6] MR Sabuncu, BTT Yeo, K Van Leemput, B Fischl, and P Golland, "A generative model for image segmentation based on label fusion," *IEEE Trans. Med. Im.*, vol. 29, no. 10, pp. 1714–1729, 2010.

[7] J Ashburner and KJ Friston, "Unified segmentation," *Neuroimage*, vol. 26, no. 3, pp. 839–851, 2005.

[8] K Pohl, J Fisher, M Shenton, R McCarley, W Grimson, R Kikinis, and W Wells, "Logarithm odds maps for shape representation," *MICCAI 2006*, pp. 955–963.

[9] VS Caviness Jr, PA Filipek, and DN Kennedy, "Magnetic resonance technology in human brain science: blueprint for a program based upon morphometry.," *Brain Dev.*, vol. 11, no. 1, pp. 1–13, 1989.

[10] "Freesurfer wiki," http://surfer.nmr.mgh.harvard.edu.