Probabilistic label fusion with a parametric generative model

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 Applied Mathematics and Computer Science, DTU, Denmark
[3] Departments of Information and Computer Science and of Biomedical Engineering and Computational Science, Aalto University, Finland

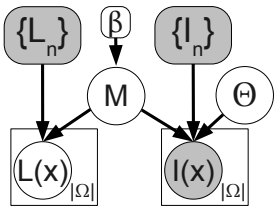
Abstract. We present a modification of our previously presented generative model for cross-modality label fusion to analyze the data from the SATA segmentation challenge with standardized registrations. We generalize our framework to multimodal data, and assume that the intensities of the atlases and the target are related by a low-order polynomial function. We present the results in terms of Dice score and compare them with the output from majority voting, showing statistically significant improvement in 20 of 22 structures of interest in the three datasets.

1 Methods

1.1 Generative model

This paper describes a label fusion algorithm to segment the three datasets (brain, cardiac, canine) in the SATA challenge using standardized, precomputed registrations. The proposed algorithm builds on the semi-parametric model we described in [1], which focused on fusion across modalities, i.e., when the atlases and the test scan to segment have been acquired with different modalities or MRI sequences. In the SATA challenge, this is not the case. Therefore, we modified the algorithm to exploit the fact that the atlases and test scan have similar intensity profiles.

The proposed method is based on the generative model shown in Fig. 1a, and the corresponding equations in Fig. 1b. $\{I_n\}$ and $\{L_n\}$ correspond to the intensities and labels of the N_{atl} deformed atlases. $M(\mathbf{x}) \in \{1, \ldots, N_{atl}\}$ indexes which atlas generated the label and intensity of the voxel at spatial location $\mathbf{x}$. The spatial smoothness of M in ensured by a Markov Random Field (MRF) prior with constant β (Eq. 1 in Fig. 1b, where $\mathcal{V}_x$ is the 6-neighborhood of $\mathbf{x}$ and $\delta(\cdot)$ is Kronecker's delta). Given $M(\mathbf{x})$ for a voxel, its label $L(\mathbf{x})$ is sampled from a categorical distribution given by a logOdds model derived from the warped labels of atlas $M(\mathbf{x})$ (Eq. 2 in Fig. 1b, where ρ is the slope of the model and D_n^l is the signed distance transform for atlas n and label $l \in \{1, \ldots, \mathcal{L}\}$). The intensity $\mathbf{I}(\mathbf{x})$ follows a multivariate Gaussian where the mean is obtained by transforming the intensity of atlas $M(\mathbf{x})$ at that location (Eq. 3 in Fig. 1b). The intensity transform Γ is defined by a set of low-order polynomials (one per



(a)

1. $M \sim \frac{1}{Z(\beta)} \prod_{\mathbf{x}} \exp\left(\beta \sum_{\mathbf{y} \in \mathcal{V}_{\mathbf{x}}} \delta(M(\mathbf{x}) = M(\mathbf{y}))\right)$
2. $L(\mathbf{x}) \sim \exp\left(\rho D^{L(\mathbf{x})}_{M(\mathbf{x})}(\mathbf{x})\right) / \sum_{l=1}^{\mathcal{L}} \exp\left(\rho D^{l}_{M(\mathbf{x})}(\mathbf{x})\right)$
3. $\mathbf{I}(\mathbf{x}) \sim p(\mathbf{I}(\mathbf{x})|\mathbf{I}_{M(\mathbf{x})}(\mathbf{x}), \theta) = \mathcal{N}\left(\mathbf{I}(\mathbf{x}); \Gamma(\mathbf{I}_{M(\mathbf{x})}; \mathbf{C}), \mathbf{\Sigma}\right)$

(b)

Fig. 1. Graphical model (a) and corresponding equations (b). Random variables are in circles, constants in boxes, observed variables shaded and plates indicate replication.

image channel) with coefficients $\mathbf{C} = [\mathbf{d}_1, \ldots, \mathbf{d}_{N_{chan}}]$. We group the covariance of the Gaussian $\mathbf{\Sigma}$ and the coefficients of the polynomials into a vector of model parameters $\theta = (\mathbf{C}, \mathbf{\Sigma})$. A flat prior on Θ, i.e. $p(\Theta) \propto 1$, completes the model.

In this paper, we make Γ a second order polynomial. In [1] we used a semi-parametric model in which Γ does not need to be continuous; while this setup can handle fusion across modalities, we believe that the polynomial is a more appropriate choice for the intramodality label fusion scenarios in the SATA challenge. A key assumption of the model is that the intensities are consistent across atlases, which requires some preprocessing.

1.2 Inference

The segmentation $\hat{L}$ of image I is computed by maximizing $p(L|I, \{I_n\}, \{L_n\})$. Following [1], it can be shown that the approximate solution can be iteratively computed with the following variational expectation maximization algorithm:

1. Initialize $q_x(m) = 1/N_{atl}$, $m = 1, \ldots, N_{atl}$; this is a categorical distribution of atlas indices at each voxel.
2. E-step: update q_x with fixed points iterations of:

$$q'_x(m) = p(\mathbf{I}(\mathbf{x})|\mathbf{I}_m(\mathbf{x}), \theta) \exp[\beta \sum_{y \in \mathcal{V}_x} q_y(m)], \quad q_x(m) = q'_x(m) / \sum_{m=1}^{N_{atl}} q'_x(m).$$

3. M-step (part 1): update $\mathbf{\Sigma} = \frac{\sum_{\mathbf{x}} \sum_{m=1}^{N_{atl}} q_x(m)[\mathbf{I}(\mathbf{x}) - \Gamma(\mathbf{I}_m(\mathbf{x}); \mathbf{C})][\mathbf{I}(\mathbf{x}) - \Gamma(\mathbf{I}_m(\mathbf{x}); \mathbf{C})]^t)}{N_{voxels}}$.
4. M-step (part 2): update the polynomial coefficients $\mathbf{C}$ doing gradient ascent on: $\sum_{\mathbf{x}} \sum_{m=1}^{N_{atl}} q_x(m) \log \mathcal{N}\left(\mathbf{I}(\mathbf{x}); \Gamma(\mathbf{I}_{M(\mathbf{x})}; \mathbf{C}), \mathbf{\Sigma}\right)$
5. If not converged, go back to E-step.
6. Compute the final segmentation as $\hat{L}(\mathbf{x}) \approx \text{argmax}_{L(\mathbf{x})} \sum_m q_x(m) p(L(\mathbf{x})|L_m)$.

2 Experiments and results

The training scans in each dataset were segmented in a leave-one-out fashion. The intensities of the scans playing the role of atlases were matched to each

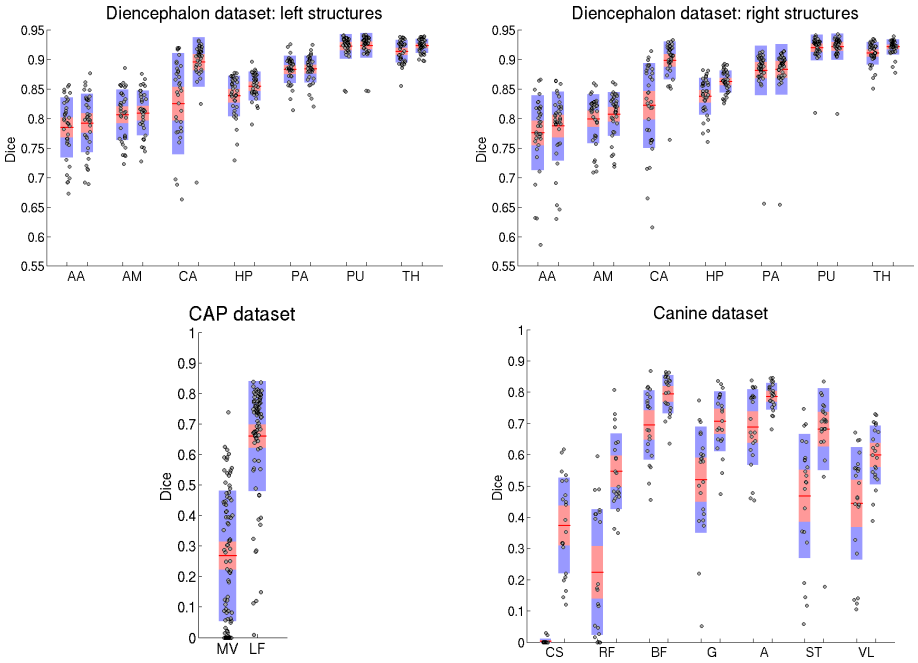


Fig. 2. Boxplots of the Dice scores. For each structure, the plot on the left corresponds to majority voting (MV), and the one on the right to label fusion (LF). The bands mark the 95% confidence interval of the mean (red) and one standard deviation (blue). AA = accumbens, AM = amygdala, CA = caudate, HP = hippocampus, PA = pallidum, PU = putamen, TH = thalamus, CS = cranial sartorius, RF = rectus femoris, BF = biceps femoris, G = gracilis, A = adductor, ST = semitendinosus, VL = vastus lateralis.

other as follows. For each atlas, the median intensity within each label (other than the background) was first computed. Then the median of these medians was taken, and the atlas intensities divided by the resulting value.

Baseline segmentations were first obtained through majority voting [2] with soft probabilities derived from the LogOdds model. The output was binarized and dilated with a spherical kernel (rad. 20 mm) to define a mask that was used in the fusion, limiting the computational requirements of the algorithm and minimizing the impact of the bias field. The parameters were set to $\rho = \beta = 1.0$. For the cardiac and canine datasets, since some registrations are very poor, we only fused the 10 best warped atlases, ranked with the sum of squared differences.

Boxplots for the Dice scores corresponding to each structure and dataset are shown in Fig. 2. Tables 1, 2 and 3 show, for the canine and brain datasets, the means and standard deviations, as well as p-values (paired Wilcoxon) comparing the two competing methods. For the cardiac dataset, the Dice scores were 26.7±21.4 (majority voting), 65.9±18.1 (label fusion), p=2.5e-15.

Method	AA	AM	CA	HP	PA	PU	TH
Maj.Vot.	78.5±5.1	80.6±4.3	82.5±8.6	83.8±3.5	88.3±2.3	92.3±2.1	91.4±1.9
Lab.Fus.	79.2±5.0	80.9±3.8	89.6±4.2	85.5±2.4	88.4±2.3	92.4±2.1	92.3±1.2
p value	3.1e-4	6.9e-2	2.7e-7	4.5e-7	1.9e-1	1.3e-4	7.2e-6

Table 1. Mean Dice scores and p-values for non-parametric paired test (brain, left).

Method	AA	AM	CA	HP	PA	PU	TH
Maj.Vot.	77.6±6.3	80.0±4.1	82.2±7.2	83.8±3.1	88.2±4.2	92.0±2.2	91.1±1.9
Lab.Fus.	78.7±5.8	80.7±3.7	89.9±3.2	86.3±1.9	88.3±4.3	92.2±2.2	92.2±1.3
p value	6.9e-5	4.0e-4	2.5e-7	2.5e-7	4.6e-3	1.2e-6	3.5e-7

Table 2. Mean Dice scores and p-values for non-parametric paired test (brain, right).

Method	CS	RF	BF	G	A	ST	VL
Maj.Vot.	0.2±0.8	22.4±20.1	69.4±11.2	51.9±17.0	68.8±12.1	46.8±19.9	44.4±18.0
Lab.Fus.	37.3±15.3	54.6±12.1	79.4±6.2	70.7±9.6	78.6±4.3	68.1±13.2	59.9±9.4
p value	4.0e-5	6.1e-5	2.3e-4	8.0e-5	1.2e-3	4.6e-5	2.4e-3

Table 3. Mean Dice scores and p-values for non-parametric paired test (canine).

3 Discussion

The results show consistent improvements of label fusion over majority voting; the Dice scores are significantly higher for all structures but the left amygdala and left pallidum. For the brain dataset, the registrations are good and majority voting already gives very good results; the improvement from label fusion is small. In the cardiac and canine datasets, the registrations are poorer, and label fusion gives a considerable improvement over majority voting. In these datasets, the Dice scores are low in absolute terms; if the accuracy of the alignments cannot be improved, non-local techniques would have to be used to achieve good segmentation results through label fusion.

Acknowledgements

This research was supported by NIH NCRR (P41-RR14075), NIBIB (R01EB006758, R01EB013565, 1K25EB013649-01), NINDS (R01NS052585), NIH 1KL2RR025757-01, Academy of Finland (133611), TEKES (ComBrain), Harvard Catalyst, and financial contributions from Harvard and affiliations.

References

1. Iglesias, J., Sabuncu, M., Van Leemput, K.: A probabilistic, non-parametric framework for inter-modality label fusion. In: Proceedings of MICCAI. (2013)
2. Rohlfing, T., Brandt, R., Menzel, R., Maurer, C.: Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. NeuroImage **21**(4) (2004) 1428–1442