Building an Ensemble of Complementary Segmentation Methods by Exploiting Probabilistic Estimates

Gerard Sanroma[1], Oualid M. Benkarim[1], Gemma Piella[1], and Miguel Ángel González Ballester[1,2]

[1] Univ. Pompeu Fabra, Barcelona, Spain
[2] ICREA, Barcelona, Spain
gerard.sanroma@upf.edu

Abstract. Two common ways of approaching atlas-based segmentation of brain MRI are (1) intensity-based modelling and (2) multi-atlas label fusion. Intensity-based methods are robust to registration errors but need distinctive image appearances. Multi-atlas label fusion can identify anatomical correspondences with faint appearance cues, but needs a reasonable registration. We propose an ensemble segmentation method that combines the complementary features of both types of approaches. Our method uses the probabilistic estimates of the base methods to compute their optimal combination weights in a spatially varying way. We also propose an intensity-based method (to be used as base method) that offers a trade-off between invariance to registration errors and dependence on distinct appearances. Results show that sacrificing invariance to registration errors (up to a certain degree) improves the performance of our intensity-based method. Our proposed ensemble method outperforms the rest of participating methods in most of the structures of the NeoBrainS12 Challenge on neonatal brain segmentation. We achieve up to $\sim 10\%$ of improvement in some structures.

Keywords: multi-atlas segmentation, ensemble learning, patch-based label fusion, brain MRI

1 Introduction

Segmentation of brain tissues and structures is important for quantitative analysis of neuroimaging data. Two major trends of approaching this problem are multi-atlas label fusion and intensity-based modelling. Multi-atlas label fusion (MALF) consists of registering a set of atlases onto the target image, and then combining their labelmaps into a consensus target segmentation [4, 14, 10]. The label on each target point is usually computed as a weighted voting of the registered atlas labels in a neighborhood. Such weights, denoting the importance of each atlas decision, are usually computed by means of local image (i.e., patch) similarity. Patch similarity is highly effective at delimiting different structures, even when they share similar appearance characteristics. However, a reasonable

registration is needed, since anatomical correspondence is based on a neighborhood search. This is partially alleviated by using multiple atlases, but it is a problem when using only a few atlases, especially in the case of convoluted structures. Also, enlarging the neighborhood search beyond certain limits actually degrades the performance due to false positives issues with the patch similarity criteria. Another kind of methods use image intensity-based models, either generative [3] or discriminative ones [1]. Most of these methods assume a global intensity model for each structure, so they are robust to registration errors to some extent (or do not require registration at all). This is an advantage for convoluted structures which are difficult to register, but it is a problem for the ones with weak appearance cues. As our first contribution, we propose a discriminative method that alleviates this problem by breaking down the modeling of the intensities in the image into different (automatically defined) regions. By building region-specific models, we improve the discrimination based on intensity at the cost of relying on a rough spatial localization of the regions in the image. A similar idea was explored in [8], although they used generative models and manually-defined regions.

The main contribution of this paper is an ensemble segmentation method that combines the complementary features of intensity-based and label fusion approaches. Following the idea of stacking [7], we learn the systematic combination of base methods that produces the best accuracy. Our method does not require any additional data besides the atlases already used by the base methods. To further limit the dependence on the number of atlases, we use the probabilistic estimates of the base methods (instead of the crisp segmentations), which improves the generalization abilities of stacking [7]. Although we focus on combining intensity-based and label fusion methods, the methodological framework described in this paper can be used to combine any kind of base methods as long as they provide probabilistic estimates. Fig. 1 shows the pipeline of our method. A related work by Ledig *et al.* [6] proposed to refine the results of label fusion by using intensity-based models. Such refinement was based on the variation of intensity distributions between the refined and non-refined segmentations. This heuristic is suitable for correcting errors that produce large variations in the intensity distributions (typically, merging csf-like with non-csf-like structures), whereas our method deals with learning a systematic combination of an arbitrary set of methods which is optimal given the set of available atlases.

2 Method

We denote the to-be-segmented target image as T, with T_i denoting the image intensity at voxel i. The atlases consist of a set of n images and labelmaps, where A_{ij} and $L_{ij} \in \{1, \ldots, p\}$ denote, respectively, the intensity value and anatomical structure present at voxel i of j-th atlas (we assume that atlas images and labelmaps are registered to the target). We first present our regional learning-based method, as a representative of intensity-based methods, and then we move on to present our proposed ensemble method.

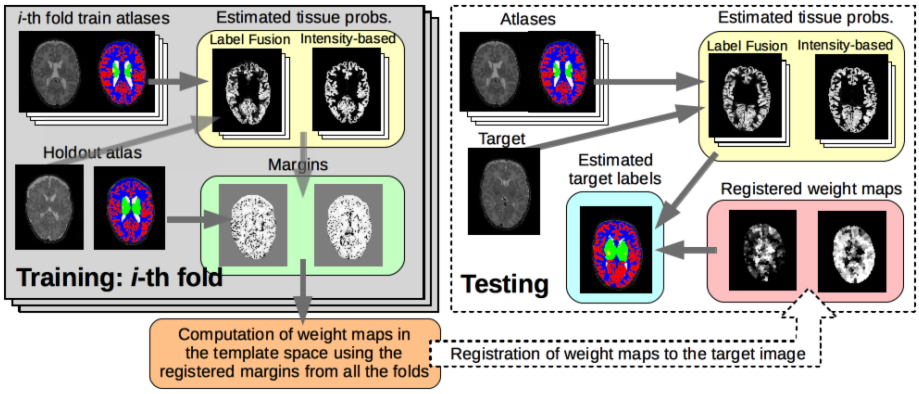

Fig. 1. Pipeline of the method.

2.1 Regional Learning-based Segmentation

Learning-based methods aim at finding a function $f : \mathbb{R}^d \rightarrow \{1, \ldots, p\}$ mapping a set of image features with their corresponding anatomical label. In supervised learning, we use a training set drawn from the atlases, denoted as $\{\mathbf{x}_{ij}, y_{ij} | i \in \Omega\}_{j=1}^{n}$, using the whole image domain Ω, where $\mathbf{x}_{ij}$ and y_{ij} denote the image features and anatomical label from atlas j at position i, respectively. Note that this approach is invariant to registration, since the classifier is learned on the whole image domain, but it assumes stable appearance properties across the image. An option to roughly take into account the spatial information is to include the position i into the feature vector as done by [1]. Instead, we propose to partition the image into disjoint parcels, denoted as $\{\Gamma_r | \bigcup_r \Gamma_r = \Omega\}$, and learn a different classifier f_r for each region. In order to generate appropriate regions for the classifiers, we use the SLIC [11] algorithm to divide a template image into super-voxels. Note that we need the training images to be registered to a template in order to draw the samples from each region. Our goal is to learn a classifier f_r for each region Γ_r independently. In the case of SVM, we compute each regional classifier f_r by optimizing the following objective:

$$\min_{f_r} \quad \mathrm{reg}\,(f_r) + C \sum_{j}^{n} \sum_{i \in \Gamma_r} \xi_{ij} \quad \text{s.t.} \quad y_{ij} f_r\big(\mathbf{x}_{ij}\big) \geq 1 - \xi_{ij} \tag{1}$$

where $\mathrm{reg}\,(\cdot)$ is the regularization term penalizing highly complex functions, ξ_{ij} indicates the error incurred by each training sample and C controls the trade-off between data fitting and regularization.

In the testing phase, the (crisp) label at position $i \in \Gamma_r$ on a new target image, denoted as F_i, is assigned using the corresponding regional classifier $F_i = f_r\big(\mathbf{x}_i\big)$, where $\mathbf{x}_i$ denotes the target image features at position i. In order to obtain probabilistic estimates to be used by our ensemble method, we apply the method by Wu *et. al.* [16].

2.2 Ensemble Segmentation based on Probabilistic Estimates

Based on the observation that label fusion and intensity-based segmentation methods have complementary features, we aim at finding the optimal combination at the different regions. We follow the idea of stacking which consists in combining the prediction of the base methods, in particular their probabilistic estimates, in an effective way. Formally, the target label at position i is computed as follows:

$$F_i = \arg \max_{s \in \{1,\dots,p\}} \sum_k \omega_{ik} P_{is}^k \quad , \tag{2}$$

where P_{is}^k is the probability of voxel i having label s according to the k-th segmentation method (yellow panels in Fig. 1) and ω_{ik} is the weight for method k at position i (red panels in Fig. 1).

The margin of a classifier is related to the distance of the samples to the classification boundary. The higher the value of the margin, the lower the risk of misclassification. The margin of a base method at point i can be defined as $m_i^k = \Lambda_{ik} P_{ic}^k$, where $\Lambda_{ik} \in \{1, -1\}$ denotes whether the predicted label by method k (say, c) is correct (1) or not (-1) and P_{ic}^k is the confidence of such prediction. The margin is positive in case of correct prediction and negative otherwise (proportionally to the confidence of the prediction). The green panels in Fig. 1 show the margins of the base methods. Here, the notion of complementarity is nicely captured by the fact that their margins should be uncorrelated.

We compute the weights of the ensemble in a leave-one-atlas-out fashion (gray panel in Fig. 1). That is, we use each method to segment the hold-out atlas using the rest of the atlases (yellow panel in Fig. 1). For each point i, the margin of the ensemble is defined as [7]:

$$m_i(\mathbf{w}) = \sum_k \omega_k \Lambda_{ik} P_{ic}^k \quad \text{s.t.} \quad \sum_k \omega_k = 1 \quad , \tag{3}$$

where $\omega_k \in \mathbf{w}$ is the weight of method k and $\Lambda_{ik} P_{ic}^k$ is its margin. To avoid the loss of precision, the margins of each method are computed in the native space of each hold-out atlas and then transformed to the template space to compute the weights (orange panel in Fig. 1). Instead of computing a weights $\mathbf{w}$ for each point, we use a certain step size and compute the weights for a neighborhood $\mathcal{N}$. We compute the weights that minimize the following quadratic loss:

$$\min_{\mathbf{w}} \sum_{i \in \mathcal{N}} (1 - m_i(\mathbf{w}))^2 + \lambda \|\mathbf{w}\|^2 = \min_{\mathbf{w}} \|U - M\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_2^2 \quad , \tag{4}$$

where U is a vector of ones, M is a matrix with each column M_k containing the margins of method k for all the neighborhood (i.e., $M_k = \left[m_1^k, \dots, m_i^k, \dots, m_{|\mathcal{N}|}^k\right]^\top$, $i \in \mathcal{N}$) and λ is a regularization parameter.

The computed weights can be used to segment any new target image using the same atlases and segmentation methods as used for training. Given a new target image T, first, the segmentation probabilities are obtained by each of the

methods (yellow panel in Fig. 1). Then, the weight maps are registered to the target space and the target labels are obtained as denoted in eq. (2) (blue panel in Fig. 1).

3 Experiments

We present tissue segmentation experiments in the IBSR dataset [15] and tissue / sub-cortical structure segmentation experiments in the NeoBrainS12 challenge dataset [5]. Prior to segmentation we correct inhomogeneities with the N4 algorithm [13] and match the histograms to a template image [9]. Template images have been built from the images of the respective datasets using the ANTs [2] script `buildtemplateparallel.sh`. Images have been non-rigidly registered to the template, and pairwise atlas-target registrations required by the base methods are obtained by concatenating the registrations through the template.

3.1 IBSR dataset

The IBSR dataset [15] contains 18 images with manual annotations of several tissues. We evaluate the segmentation performance on white matter (WM), gray matter (GM) and ventricular cerebrospinal fluid (CSF). We use 4 images as atlases and the remaining 14 as targets. We try to select 4 representative atlases according to their distribution in the manifold.

We combine joint label fusion (JointLF) with the regional learning-based (RegLB) method proposed in Sec. 2.1. For RegLB, we use the following image features: image intensity, laplacian and magnitude of the gradient after convolution with a Gaussian kernels with $\sigma = 1, 2, 3$ (we did not find any improvement by including position as feature). We randomly pick 5% of the data to train the classifier for each region.

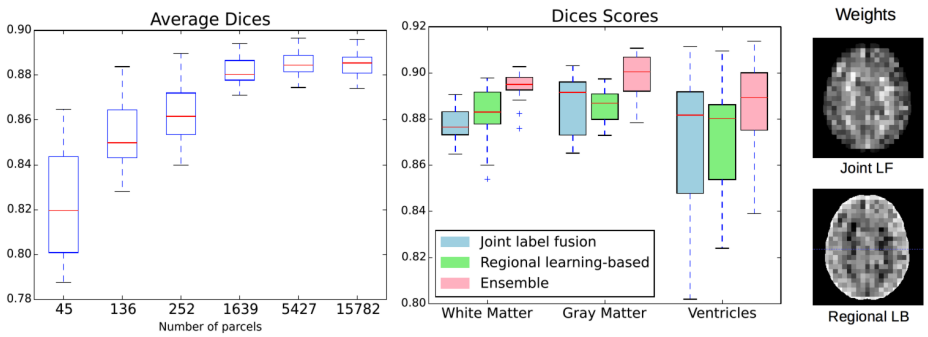

Fig. 2. Results in the IBSR dataset.

First, in Fig. 2 (left) we show the boxplots of the Dice scores obtained by increasing the number of parcels in RegLB. As we can see, the accuracy increases

by decreasing the size of the parcels until the top performance is reached at ~ 5000 parcels, thus confirming the hypothesis that better results are obtained with more local models. In Fig. 2 (middle) we show the boxplots of the Dice scores by JointLF, RegLB (5000 parcels) and their combination as described in Sec. 2.2. As we can see, results by JointLF and RegLB are similar in terms of Dice score. However, their combination improves considerably, thus confirming the advantages of combining complementary methods (we also tried to combine different RegLB at multiple parcellation levels, obtaining no improvement). Finally, Fig. 2 (right) shows a slice of the weight maps obtained for each of the base methods. As we can see, RegLB has higher weights in the cortical area (a highly convoluted area) whereas JointLF has higher weights in the interior.

3.2 NeoBrainS12 Challenge

The NeoBrainS12 Challenge dataset contains T1 and T2 scans of neonates at 30 weeks and 40 weeks gestational age (GA), divided into training and testing sets. Training images include manual annotations of the following 8 structures: cortical gray matter, basal ganglia and thalami, unmyelinated white matter, myelinated white matter, brainstem, cerebellum, ventricles and cerebrospinal fluid in the extracerebral space. Participants have no access to the manual annotations on the testing set. There are 2 training and 5 testing images at 30 weeks coronal, 2 training and 5 testing images at 40 weeks axial, and 5 remaining testing images at 40 weeks coronal. We use the 2 training images at each gestational age as atlases in our method. Here, we combine joint label fusion [14] (JointLF) and the intensity-based method Atropos [3]. We use the probabilistic results of JointLF as spatial prior for Atropos. There is a parameter $0 \leq w \leq 1$ regulating the importance of the prior in Atropos, with $w = 0$ resulting in segmentations totally driven by image intensities and $w = 1$ resulting in segmentations very similar to JointLF. The optimal value of this parameter, based on the results in the training set, were $w = 0.2$ for 30 weeks GA and $w = 0.1$ for 40 weeks GA. In Table 1 we show the average Dice scores obtained by our method over the testing images (as reported by the organizers[3]) in each of the structures, along with the best score obtained among *all* the rest of methods.

As we can see, our method obtains the best results in the majority of the structures overall. It is worth noting that we are comparing our results with the best performing method among the rest, which is usually different for each structure. Our method achieves improvements of up to $\sim 10\%$ in some structures such as the brainstem and the challenging myelinated white matter (MWM). Fig. 3 shows some qualitative segmentation results on the testing images for 30 weeks (top) and 40 weeks (bottom) gestational age. Yellow ellipses denote interior structures where Atropos fails likely due to weak appearance cues. Blue ellipses denote cortical zones where JointLF fails likely due to registration artifacts. Our ensemble method obtains a satisfactory combination in all areas.

[3] Results should appear as "anonymous" at http://neobrains12.isi.uu.nl/mainResults.php

30 weeks coronal										
	Cb	MWM	BGT	Vent	UWM	BS	CoGM	CSF	U+M	CSF+V
Ensemble	**0.92**	0.66	**0.90**	**0.88**	**0.93**	**0.86**	**0.75**	**0.85**	**0.93**	**0.86**
Best score*	0.88	**0.69**	0.84	**0.88**	0.91	0.76	0.71	0.83	0.90	0.84
40 weeks axial										
	Cb	MWM	BGT	Vent	UWM	BS	CoGM	CSF	U+M	CSF+V
Ensemble	**0.94**	**0.54**	**0.93**	0.83	**0.91**	**0.85**	0.85	**0.79**	0.90	0.79
Best score*	0.92	0.47	0.92	**0.86**	0.90	0.83	**0.86**	**0.79**	**0.92**	**0.80**
40 weeks coronal										
	Cb	MWM	BGT	Vent	UWM	BS	CoGM	CSF	U+M	CSF+V
Ensemble	0.91	0.33	**0.89**	**0.85**	0.87	**0.76**	0.73	0.72	**0.86**	0.73
Best score*	**0.92**	**0.48**	0.88	0.84	**0.89**	0.75	**0.77**	**0.77**	0.84	**0.79**

Table 1. Dice scores in the NeoBrainS12 Challenge. ($*$ Best score obtained among the rest of participants.)

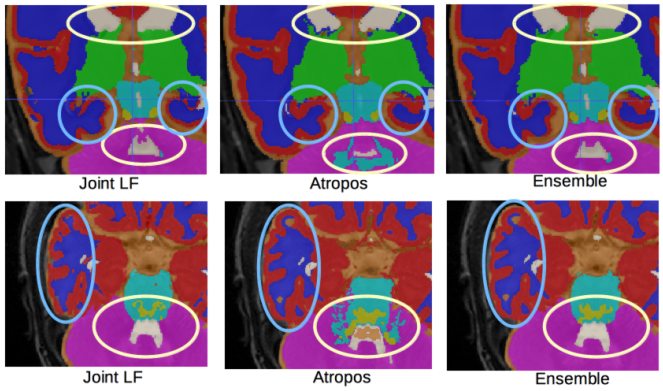

Fig. 3. Qualitative results, for 30 weeks (top) and 40 weeks (bottom) GA.

4 Conclusions

We presented an ensemble segmentation method to combine label fusion and intensity-based segmentation that uses probabilistic estimates of the base methods. We also proposed an intensity-based method with arbitrarily defined spatial domain. Results show that reducing the spatial domain of the intensity-based models considerably improves the segmentation performance. Results also show that the combination of complementary methods, such as label fusion and intensity-based, using the proposed ensemble framework considerably improves the results. We outperform the rest of the methods in the segmentation of the majority of the structures in neonatal images from the NeoBrainS12 Challenge.

Acknowledgements The first author is co-financed by the Marie Curie FP7-PEOPLE-2012-COFUND 462 Action. Grant agreement no: 600387

References

1. Anbeek, P., Isgum, I., van Kooij, B.J.M., Mol, C.P., Kersbergen, K.J., Groenendaal, F., Viergever, M.A., de Vries, L.S., Benders, M.J.N.L.: Automatic segmentation of eight tissue classes in neonatal brain mri. PLOS ONE 8(12) (2013)
2. Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C.: Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis 12(1), 26–41 (2008)
3. Avants, B.B., Tustison, N.J., Wu, J., Cook, P.A., Gee, J.C.: An open source multivariate framework for n-tissue segmentation with evaluation on public data. Neuroinformatics 9(4), 381–400 (2011)
4. Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L.: Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. NeuroImage 54(2), 940–954 (2011)
5. Isgum, I., Benders, M.J.N.L., Avants, B., Cardoso, M.J., Counsell, S.J., Gomez, E.F., Gui, L., Hüppi, P.S., Kersbergen, K.J., Makropoulos, A., Melbourne, A., Moeskops, P., Mol, C.P., Kuklisova-Murgasova, M., Rueckert, D., Schnabel, J.A., Srhoj-Egekher, V., Wu, J., Wang, S., de Vries, L.S., Viergever, M.A.: Evaluation of automatic neonatal brain segmentation algorithms: the neobrains12 challenge. Medical Image Analysis 20(1), 135–151 (2015)
6. Ledig, C., Heckemann, R.A., Hammers, A., Lopez, J.C., Newcombe, V.F.J., Makropoulos, A., Lötjönen, J., Menon, D.K., Rueckert, D.: Robust whole-brain segmentation: Application to traumatic brain injury. Medical Image Analysis 21, 40–58 (2015)
7. Li, L., Hu, Q., Wu, X., Yu, D.: Exploration of classification confidence in ensemble learning. Pattern Recognition 47, 3120–3131 (2014)
8. Makropoulos, A., Gousias, I.S., Ledig, C., Aljabar, P., Serag, A., Hajnal, J.H., Edwards, A.D., Counsell, S.J., Rueckert, D.: Automatic whole brain mri segmentation of the developing neonatal brain. IEEE TMI 33(9), 1818–1831 (2014)
9. Nyúl, L.G., Udupa, J.K.: On standardizing the mr image instensity scale. Magnetic Resonance in Medicine 42(6), 1072–1081 (1999)
10. Sanroma, G., Benkarim, O.M., Piella, G., Wu, G., Zhu, X., Shen, D., González-Ballester, M.A.: Discriminative dimensionality reduction for patch-based label fusion. In: Machine Learning Meets Medical Imaging (2015)
11. Shaji, R.A.A., Smith, K., Lucchi, A., Fua, P., Susstrünk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(11), 2274–2282 (2012)
12. Ting, K.M., Witten, I.H.: Issues in stacked generalization. Journal of Artificial Intelligence Research 10, 271–289 (1999)
13. Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C.: N4itk: Improved n3 bias correction. IEEE Transactions on Medical Imaging 29(6), 1310–1320 (2010)
14. Wang, H., Suh, J.W., Das, S.R., Pluta, J.B., Craige, C., Yushkevich, P.A.: Multi-atlas segmentation with joint label fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(3), 611–623 (2013)
15. Worth, A.J.: The internet brain segmentation repository (ibsr)
16. Wu, T.F., Lin, C.J., Weng, R.C.: Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research 5, 975–1005 (2004)