Enhanced Probabilistic Label Fusion by Estimating Label Confidences Through Discriminative Learning

Oualid M. Benkarim[1(✉)], Gemma Piella[1], Miguel Angel González Ballester[1,2], and Gerard Sanroma[1]

[1] Universitat Pompeu Fabra, Barcelona, Spain
oualid.benkarim@upf.edu
[2] ICREA, Barcelona, Spain

Abstract. Multiple-atlas segmentation has recently shown success in automatic segmentation of brain images. It consists in registering the labelmaps from a set of atlases to the anatomy of a target image, and then fusing the multiple labelmaps into a consensus segmentation on the target image. Accurately estimating the confidence of each atlas decision is key for the success of label fusion. Common approaches either rely on local patch similarity, probabilistic statistical frameworks or a combination of both. We present a probabilistic label fusion framework that takes into account label confidence at each point. Maximum likelihood atlas confidences are estimated by explicitly modelling the relationship between image appearance and segmentation errors. We also propose a novel type of label-dependent appearance features based on atlas labelmaps. Our results indicate that the proposed label fusion framework achieves state-of-the-art performance in the segmentation of subcortical structures.

Keywords: Multiatlas segmentation · Confidence estimation · Discriminative learning · brain MRI

1 Introduction

Multiple-atlas segmentation has shown to be a promising technique for brain structural segmentation [3]. It consists in propagating the labelmaps from a set of atlases to a target image. There are two main steps: (1) image registration, where the spatial transformations are computed to warp the atlas labelmaps to the target image, and (2) label fusion, where these candidate segmentations are fused into a consensus segmentation. When using multiple atlases rather than a single atlas, we adapt better to the anatomical variability in the target image. The label fusion problem consists in defining the optimal combination of atlases at each region of the target image. The simplest approach, known as majority voting (MV) [7], assigns each target voxel the most frequent label occurring among the atlases. This method has shown promising results, however, since all

S. Ourselin et al. (Eds.): MICCAI 2016, Part II, LNCS 9901, pp. 505–512, 2016.
DOI: 10.1007/978-3-319-46723-8_58

the atlases are combined with equal weight, having atlases too dissimilar to the target will push the resulting segmentation away from the true target anatomy.

A more reasonable approach would require the definition of a *confidence* measure for each atlas reflecting their reliability in segmenting the target image and giving more weight during the combination to those atlases with higher confidence. A possible strategy is to assign each registered atlas a global weight based on its similarity with the target image [1]. However, image similarity sometimes does not correlate well with atlas confidence [8]. Another kind of approaches alleviate this problem by estimating the atlas confidence by using a more direct measure of the anatomical overlap [10]. These approaches alternate the segmentation of the target anatomy and the estimation of the atlas confidences in an iterative fashion. As an example of these methods, STAPLE [10] defines a principled statistical framework to perform such estimation. However, these methods do not take into consideration intensity information available from the images.

The so-called patch-based label fusion methods (PBLF) estimate local confidences of each atlas for each target point based on local image similarity [5,9]. Joint label fusion [9], for instance, models pairwise dependency between atlases to reduce the weights of correlated atlases. Other approaches incorporate the versatility of local similarity-based approaches into the framework of STAPLE. For example, STEPS [4] improves the target anatomy estimation in STAPLE by using a methodology inspired by PBLF. As pointed out earlier, using image similarity can induce a bias in the estimation of the confidence.

We propose a probabilistic label fusion framework that takes into account local atlas confidences at each point. In the training phase, for each atlas, we compute their confidence models by maximum likelihood estimation. To do so, we register the rest of the atlases to each different atlas space. Confidence estimation in the space of each atlas is important in order to deal with systematic segmentation errors caused by registration failures. We propose two ways of estimating the confidence models: (1) a simple one depending on local label statistics, and (2) an advanced one modelling the relationship between local image appearance and segmentation errors. In the testing phase, *spatial confidence maps* (SCMs) are obtained for a given target image using the confidence models computed in the training phase. Target labels are then estimated with the proposed framework using the SCMs in conjunction with the atlas labelmaps. Furthermore, we propose a new feature extraction process that takes into account the atlas labels. Figure 1 shows the pipeline of the method.

The outline of the paper is as follows. In Sect. 2 we present the details of our method. In Sect. 3 we describe the experimental setting and present the results, and in Sect. 4 we conclude the paper.

2 Method

2.1 Enhanced Probabilistic Label Fusion

Consider we have a target image T, where T_i denotes the intensity value at voxel i, and a set of atlas images $\mathbf{A}$ along with their labelmaps $\mathbf{D}$, where $D_{ij} \in \mathbf{D}$ and

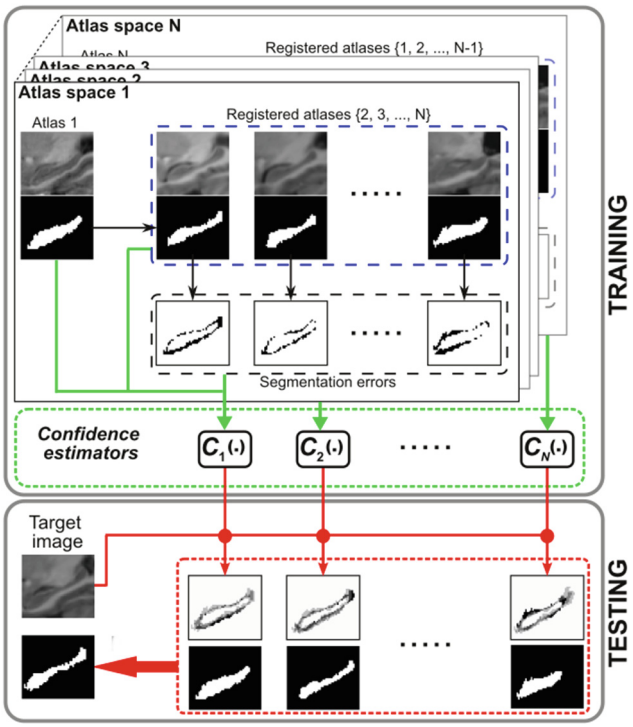


Fig. 1. Pipeline of the proposed method. **Training**: the remaining atlases are registered onto each atlas and the confidence models are computed. **Testing**: given the new image, the SCMs are obtained using the confidence models. Target labels are then estimated according to the proposed label fusion framework.

$D_{ij} = \{1, \ldots, p\}$, indicates which one of the p structures is present at voxel i of the j-th atlas. We denote the to-be-estimated target labelmaps as F.

We want to find the target labels that maximize the following posterior probability:

$$f\left(F|\mathbf{D}, \mathbf{C}\right) = \prod_i f\left(F_i|\mathbf{D}_i, \mathbf{C}_i\right) = \prod_i \frac{f\left(\mathbf{D}_i|F_i, \mathbf{C}_i\right) f\left(F_i\right)}{f\left(\mathbf{D}_i, |\mathbf{C}_i\right)}, \tag{1}$$

where $\mathbf{D}_i$ denotes the set of atlas decisions for voxel i and $\mathbf{C}_i$ denotes their respective confidences. Note that we assume conditional independence in the target voxels. Further assuming independence among the atlas decisions we obtain:

$$f\left(F_i|\mathbf{D}_i, \mathbf{C}_i\right) = \frac{\prod_j f\left(D_{ij}|F_i, C_{ij}\right) f\left(F_i\right)}{\sum_{s\in\{1,0\}} \prod_j f\left(D_{ij}, |F_i = s, C_{ij}\right) f\left(F_i = s\right)}. \tag{2}$$

Here we assume that we have only two labels denoted $\{0, 1\}$. The case of multiple labels can be handled in a one-versus-rest fashion.

Accordingly, the probability of the target label F_i being foreground (i.e., 1) is defined as:

$$f\left(F_i = 1|\mathbf{D}_i, \mathbf{C}_i\right) = \frac{a_i}{a_i + b_i}, \tag{3}$$

where $a_i = f(F_i=1) \prod_j f(D_{ij}|F_i=1, C_{ij})$ and $b_i = f(F_i=0) \prod_j f(D_{ij}|F_i=0, C_{ij})$.

Here, the important quantity is $f\left(D_{ij}|F_i = s, C_{ij}\right)$, which is the probability of observing decision of j-th atlas on voxel i, given that the target label is s and the atlas confidence at that point is C_{ij}. This term expresses the likelihood that the atlas and target labels coincide, and is defined as:

$$f\left(D_{ij}|F_i = s, C_{ij}\right) = \begin{cases} C_{ij} & \text{if } \; D_{ij} = s \\ 1 - C_{ij} & \text{otherwise.} \end{cases} \tag{4}$$

The central part in our work is the computation of the confidences C_{ij}. STAPLE-based methods compute it using tentative estimations of the target labels in an iterative online estimation of target labels and confidence parameters. On the contrary, we use a training set of target labels to compute it in an offline manner. This allows us to perform label fusion in a direct (non-iterative) way. We also estimate local confidence values for each voxel.

Let us focus on the computation of the confidence for a single voxel i of a single atlas j, denoted as $c \equiv C_{ij}$ for brevity (the same procedure is repeated for the rest of the voxels on the rest of atlases). Similarly, let us denote as $d \equiv D_{ij}$ the label at voxel i in the j-th atlas. We denote as $\mathbf{f} = \left\{\tilde{D}_{ik}, k \neq j\right\}$, the training set of target observations for the voxel i in the j-th atlas composed of the registered labelmaps of the rest of atlases. This is indicated by the blue panel in Fig. 1. We compute the confidence at each voxel by maximizing the following joint likelihood:

$$\hat{c} = \arg\max_c f\left(\mathbf{f}, d|c\right) = \arg\max_c \prod_k f\left(d|f_k, c\right) f\left(f_k|c\right), \tag{5}$$

where $f_k \in \mathbf{f}$. We discard the second term in the product since we assume that target labels are only affected by parameters in the presence of an atlas. Taking the logarithm and substituting the atlas likelihood term by its expression in (4) yields:

$$\hat{c} = \arg\max_c \sum_k \log f\left(d|f_k, c\right) = \arg\max_c \sum_{f_k = d} \log c + \sum_{f_k \neq d} \log\left(1 - c\right). \tag{6}$$

Taking derivatives, the optimal confidence is $c = \frac{n_h}{n_h + n_m}$, where n_h and n_m are the number of coincident target labels (hits) and different target labels (misses), respectively, from the atlas label. This defines our naive case.

Nevertheless, we further believe that local image appearances provide valuable clues for estimating this confidence. Therefore, we extend the previous naive method by substituting the constant confidence in (4) by a more complex function informed by the image appearances, as follows:

$$f\left(D_{ij}|F_i = s, \mathcal{C}_{ij}\right) = \begin{cases} \mathcal{C}_{ij}\left(\mathbf{t}_i, \mathbf{a}_{ij}\right) & \text{if } \; D_{ij} = s \\ 1 - \mathcal{C}_{ij}\left(\mathbf{t}_i, \mathbf{a}_{ij}\right) & \text{otherwise} \end{cases}, \tag{7}$$

where $\mathbf{t}_i$ and $\mathbf{a}_{ij}$ are image appearance features extracted from target image and j-th atlas around voxel i, and $\mathcal{C}_{ij}(\cdot)$ is a function denoting the confidence we have that the atlas label is correct given the target and atlas image appearances (as shown in the green panel in Fig. 1). By using image appearances, we can effectively capture the effects of registration errors on modeling such confidence. Again, our goal is to compute such function as to maximize the joint probability of each atlas observation given the training set (5). Using a similar development as in the naive case, we arrive at the following expression:

$$\hat{\mathcal{C}} = \arg\max_{\mathcal{C}} \sum_{f_k=d} \mathcal{C}(\mathbf{t}_k, \mathbf{a}) - \sum_{f_k \neq d} \mathcal{C}(\mathbf{t}_k, \mathbf{a}) \ , \tag{8}$$

where $\mathbf{t}_k$ and $\mathbf{a}$ denote the local image appearances of the k-th target training sample and atlas in the training set, respectively. This expression corresponds to the minimization of an empirical error subject to the constraint that the computed function must be a probability density function. For this we use support vector machines (SVM) with Platt's scaling [6].

In the testing stage, given a new target image T, it is first warped to each of the atlases. Then, SCMs are computed using the confidence functions of (8). Next, SCMs and the atlas labels are transformed back to the target space. Finally, we compute the label fusion using (1) (see red panel of Fig. 1).

2.2 Label-Dependent Feature Extraction

The simplest approach to represent the target features $\mathbf{t}_i$ is to use a local patch around the i-th voxel. Here we propose 2 extensions: (1) use a non-local means approach [5], and (2) use label-dependent features. Using non-local means is more robust as the confidence estimators are trained to take into account larger registration errors. The second contribution uses the j-th atlas label patch to extract label-dependent features from the target images. As illustrated in Fig. 2, given the label patch of the j-th atlas around the i-th voxel, we identify the target voxels corresponding to foreground and background regions (in the case of binary segmentation) and compute different summary statistics, namely, mean, maximum and minimum intensity, and the center of mass of each region. Finally, the difference between foreground and background features is calculated and the resulting features are appended to the intensity patch.

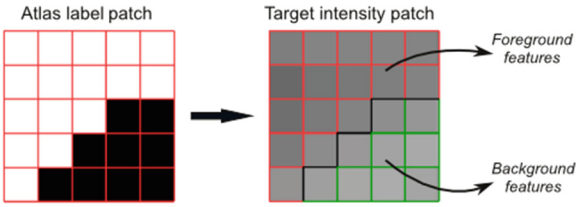


Fig. 2. Label-dependent feature extraction.

3 Experiments

In this section, we compare our approach with state-of-the-art methods in the literature for the segmentation of 7 subcortical brain structures: accumbens, amygdala, caudate, hippocampus, pallidum, putamen and thalamus proper.

3.1 Data and Preprocessing

The proposed method was evaluated on a dataset of 35 T1-weighted brain MR images from the OASIS project, with corresponding manual segmentations made public by the MICCAI 2013 challenge.[1] We registered all images to a common space using the symmetric diffeomorphic mapping (SyN) [2]. Pairwise mappings were then obtained by composing the transformation of the source image to the template and the inverse transformation of the target. Furthermore, for image intensity to be consistent across atlases, histogram matching was used.

3.2 Experimental Setup

For comparison, we considered the following methods: MV, STAPLE, STEPS and joint label fusion (JOINT). Three different versions of our method were used: (1) the naive approach, (2) SCM using only patch intensities (SCMNF), and (3) SCM with additional label-dependent features (SCMLF). Regarding the parameters, we used the default values for all methods, except for the radius of the patch and window search that was set to 1. For our method, we used SVM with a linear kernel and the penalty parameter $C = 1$. Finally, a 3-fold cross-validation procedure was used in our validation strategy. For quantitative comparison, we used the Dice similarity coefficient.

3.3 Results

Figure 3 shows a boxplot of Dice overlaps achieved by each method for all structures. There is a clear difference between the first set of methods (i.e., MV, our naive approach, STAPLE and STEPS) and the second one, including JOINT and our two SCM-based approaches. The naive method yields similar results to MV. In fact, when all atlases are used to compute the confidences, it is equivalent to MV, being the additional transformations between atlas spaces the only difference. STAPLE-based methods used in this comparison have a slightly higher Dice score than MV. Moreover, although STEPS uses image intensities to drive the fusion process, there is no consistent improvement over STAPLE, as the latter provides better overlaps in the amygdala and the pallidum.

The methods in the second set provided the most accurate segmentations, with statistically significant difference in all structures. Albeit statistically equivalent, the inclusion of label-dependent features boosted the overall performance of SCMLF compared to SCMNF. JOINT ($\mu = 0.872, \sigma = 0.049$) performed

[1] https://masi.vuse.vanderbilt.edu/workshop2013.

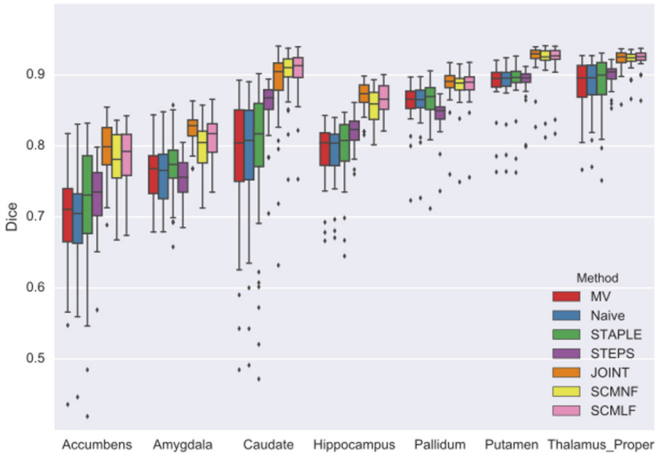
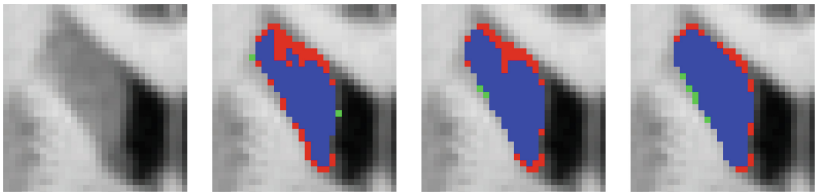


Fig. 3. Dice scores for all methods.



(a) Image (b) STEPS (c) JOINT (d) SCMLF

Fig. 4. Illustration of caudate automatic segmentation in axial view. Red and green depict manual and automatic segmentations respectively. Overlap is depicted in blue.

slightly better than SCMLF ($\mu = 0.871, \sigma = 0.054$) in terms of average overall Dice, but no statistical significance was found. Particularly, both methods yield comparable results in all structures except for the accumbens and the amygdala where JOINT provided better outcomes, and the caudate where segmentations obtained from SCMLF where more accurate as shown in Fig. 4.

Concerning computational complexity, segmentation of the accumbens, for instance, takes less than 2 s for all methods except for JOINT that takes 10 s. Our method requires an additional step for training the confidence estimators that takes around 88 min, although this is performed only once.

4 Conclusions

Registration errors are one of the main sources of systematic errors in multiatlas segmentation. In this work, we have presented a novel label fusion framework where the confidence learning process is performed in atlas space, which makes our method robust to registration errors. As opposed to STAPLE-like

approaches, SCMs estimation in our method is done offline using the available training atlases. Therefore, computational complexity at test time is comparable to the simplest approaches, such as MV. Furthermore, given the nature of the proposed method, we can include label-dependent features, supplying valuable information in the prediction of the confidences. Our experiments demonstrate that our method yields comparable results to state-of-the-art approaches.

Acknowledgments. This work is co-financed by the Marie Curie FP7-PEOPLE-2012-COFUND Action, Grant agreement no: 600387.

References

1. Artaechevarria, X., Muñoz Barrutia, A., Ortiz-de Solórzano, C.: Combination strategies in multi-atlas image segmentation: application to brain MR data. IEEE Trans. Med. Imaging **28**(8), 1266–1277 (2009)
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. Med. Image Anal. **12**(1), 26–41 (2008)
3. Babalola, K.O., Patenaude, B., Aljabar, P., Schnabel, J., Kennedy, D., Crum, W., Smith, S., Cootes, T., Jenkinson, M., Rueckert, D.: An evaluation of four automatic methods of segmenting the subcortical structures in the brain. NeuroImage **47**(4), 1435–1447 (2009)
4. Cardoso, J.M., Leung, K., Modat, M., Keihaninejad, S., Cash, D., Barnes, J., Fox, N.C., Ourselin, S.: Steps: similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation. Med. Image Anal. **17**(6), 671–684 (2013)
5. 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)
6. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in Large Margin Classifiers, pp. 61–74. MIT Press (1999)
7. Rohlfing, T., Brandt, R., Menzel, R., Maurer, C.R.: Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. NeuroImage **23**(8), 983–994 (2004)
8. Sanroma, G., Benkarim, O.M., Piella, G., Wu, G., Zhu, X., Shen, D., Ballester, M.Á.G.: Discriminative dimensionality reduction for patch-based label fusion. In: Bhatia, K.K., Lombaert, H. (eds.) MLMMI 2015. LNCS, vol. 9487, pp. 94–103. Springer, Heidelberg (2015). doi:10.1007/978-3-319-27929-9_10
9. Wang, H., Suh, J.W., Das, S.R., Pluta, J.B., Craige, C., Yushkevich, P.A.: Multi-atlas segmentation with joint label fusion. IEEE Trans. Pattern Anal. Mach. Intell. **35**(3), 611–623 (2013)
10. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging **23**(7), 903–921 (2004)