Classifier-based Multi-Atlas Label Propagation with Test-specific Atlas Weighting for Correspondence-free Scenarios

Darko Zikic[1], Ben Glocker[2], and Antonio Criminisi[1]

[1] Microsoft Research, Cambridge, UK
[2] Biomedical Image Analysis Group, Imperial College London

Abstract We propose a segmentation method which transfers the advantages of multi-atlas label propagation (MALP) to correspondence-free scenarios. MALP is a branch of segmentation approaches with attractive properties, which is currently applicable only in correspondence-based regimes such as brain labeling, which assume correspondence between atlases and test image. This precludes its use for the large class of tasks without this property, such as tumor segmentation. In this work, we propose a method which circumvents the correspondence assumption by using a classifier-based atlas representation in the spirit of the recently proposed Atlas Forests (AF). To counteract the negative effects of the over-training property of AF for applications with highly heterogeneous examples, we employ test-specific atlas weighting by the STAPLE approach. The main idea is that over-training ceases to be a problem if the prediction is based only on training atlases which are "similar" to the test image. Here, the "similarity" is based on the estimated ability of an atlas-based classifier to perform a correct labeling. We show a successful use of the proposed method for segmentation of brain tumors on data from the BraTS 2013 Challenge, which presents a correspondence-free scenario in which standard MALP cannot be expected to operate.

1 Introduction

Multi-atlas label propagation (MALP) is a popular branch of segmentation approaches. Given an atlas as a training image and the corresponding label map, the essence of MALP approaches is to perform individual atlas-based predictions, followed by a fusion step to form the final estimate. While MALP-based methods are extremely successful in certain settings such as brain labeling [1,2] or segmentation of abdominal organs [3], they are restricted to the *correspondence-based scenario*, where the assumption of correspondence between points in test and atlas images is made. In this work, we aim to transfer the advantages of the MALP framework to *correspondence-free scenarios*. Such regimes are an important class of problems in medical image analysis - they occur for example when highly heterogeneous pathologies such as tumors develop at different locations, or their shapes strongly vary. Our model problem for such a setting is multi-class segmentation of brain tumors in the BraTS 2013 challenge, which presents a challenging problem with a database of highly heterogeneous atlases [4].

Our work is motivated by two properties of the MALP framework which we aim to transfer to correspondence-free settings: 1) ability for atlas selection, and 2) computational efficiency. In the following we discuss these properties and why current MALP approaches are not applicable to correspondence-free scenarios, outline the main idea of our approach, and relate it to previous work.

A central characteristic of the MALP framework is that individual predictions are made based on each atlas, which are then fused into a final estimate. The first advantageous property resulting from the *per-atlas* characteristic is the ability for atlas selection [5]. Prediction based only on those training images which are similar to the test image has the potential to improve results, especially for underrepresented cases. This property is of increased importance for settings with highly heterogeneous atlases, and its potential can be expected to rise with the growing size of available labeled databases. The second advantageous property is the high efficiency. Recently, classifier-based MALP (CB-MALP) approaches have been introduced, which explicitly encode each atlas by an individual classifier [6,7], and significantly increase efficiency for training and experimentation compared to standard learning schemes which pool data from all atlases. The training efficiency comes from the smaller amount of samples for training of a single classifier, and the experimentation efficiency is given by the ability for cross-validation without retraining [6].

The combination of the ability for atlas selection and high efficiency makes MALP an attractive framework for general purpose segmentation, however, current MALP methods are applicable only in correspondence-based settings.

Most current MALP methods are registration-based [1,2], and thus explicitly operate in correspondence-based regimes. This holds for both, approaches based on non-linear registration which make the one-to-one correspondence assumption (e.g. [8,9,5]) , as well as for patch-based approaches which use the relaxed one-to-many assumption [10,11]. The recently proposed classifier-based MALP methods [6,7] are in principle applicable to correspondence-free scenarios. However, approaches which train a classifier on a single atlas suffer from over-training, and can be expected not to generalize well to examples very different from the training dataset. While this property is not an issue in relatively homogeneous settings such as brain labeling where above methods were shown to perform well, it becomes problematic in highly heterogeneous correspondence-free settings such as brain tumor segmentation, as our experiments confirm for the atlas forests scheme from [6]. So, despite the potential of MALP-based approaches, they are currently not used in correspondence-free scenarios. More specifically, for our model problem of brain tumor segmentation, none of the proposed methods at the BraTS challenges 2012 and 2013 [4] was set within the MALP framework.

The main idea of this work is to base the segmentation on a classifier-based MALP method, thus keeping the advantages of the MALP framework, and to counteract the over-training issues of CB-MALP schemes by using classifiers according to their ability to correctly label the test image. The rationale is that over-training becomes a smaller issue with increasing "similarity" of testing and

training data. Since the classifiers are atlas-based, such test-specific classifier weighting corresponds to weighting of training atlases based on their "similarity" to the test image. Here, the "similarity" between atlas and test image is determined by the accuracy performance of the associated atlas-based classifier, i.e. its ability to accurately label the test image. In this work, our CB-MALP method is based on the Atlas Forests (AF) framework [6], which operates by training an individual randomized forest classifier for each atlas. Originally, AF fuses the predictions by averaging the individual probabilistic classifier estimates. Instead, we propose to perform implicit atlas weighting by using the Simultaneous Truth and Performance[1] Level Estimation (STAPLE) method [12]. At test time, AF generates a set of candidate segmentations, for which STAPLE subsequently estimates the performance level, and uses these estimates as weights to combine the candidates into the final segmentation. Effectively, this means that for each test image, each of the atlases is used for prediction according to its estimated ability to perform a correct labeling. This way, our approach preserves the advantages of computational efficiency of the CB-MALP framework, while eliminating its negative effects of over-training, thus making it applicable to correspondence-free scenarios.

1.1 Relation to Prior Work

Our work is closely related to the recently proposed classifier-based MALP schemes from [6] and [7]. Our method is based on the atlas forest scheme from [6], which is an instance of CB-MALP without atlas weighting. More details of AF are discussed in Sec. 2. The focus of [7] is a generalization of STAPLE to operate on probabilistic estimates, which are in that work generated by a Gaussian Mixture Model of intensity patches, which are trained per atlas. Thus, this method is an instance of CB-MALP with implicit atlas weighting, similar to the approach proposed in this work. The studied setting in [7] is brain labeling, and application in correspondence-free scenarios is not considered.

While we are not aware of any work using a MALP-based approach for brain tumor segmentation, an interesting strategy to fuse multiple segmentations is considered in [4], where the majority vote strategy is used to fuse the results of methods which were *a priori* determined to achieve high accuracy. Since the candidate segmentations are not associated with individual atlases, this approach does not retain the MALP properties, and does not perform atlas weighting.

There are many strategies for atlas weighting for MALP. One is the use of heuristics such as intensity-based similarity of images [5,3], or subject age [5]. Heuristics are usually used to perform atlas selection (i.e. binary weighting) *prior* to testing. Alternatively, STAPLE [12] performs an implicit weighting of atlas estimates. It operates *a posteriori* on computed candidate segmentations. Its generality makes it applicable also in highly heterogeneous correspondence-free settings, for which heuristic design is difficult.

[1] In the context of STAPLE, 'performance' stands for 'accuracy', and we use the term in the same sense in this paper.

2 Method

The proposed framework consists of two steps: (1) use a classifier-based MALP method, i.e. represent an individual atlas A_i by a classifier trained only on the data from A_i, and at test, use each classifier to generate a candidate segmentation $\hat{L}_i$, and (2) perform test-specific atlas weighting based on $\{\hat{L}_i\}$. In this work, we use randomized forests (RF) for (1), and the STAPLE method [12] for (2). We briefly describe these two components below.

2.1 Atlas-based Estimates by Randomized Classification Forests

We use the general idea of the atlas forest framework [6], but modify the actual RF classifier according to [13] for the task of brain tumor segmentation. In contrast to [6], we do not incorporate any location-based features since we aim for the correspondence-free setting. Instead, as discussed in [13], we augment the multi-channel input data with class-probability estimates, and train an RF with context-aware features on this augmented data.

Given a set of N training atlases $\{A_i\}_{i=1:N}$, consisting of an intensity image I_i and the corresponding labelmap L_i, the task is to estimate a labelmap L for the test image I. As described in detail in Sec. 3.1, the original intensity images are multi-channel 3D images, with 4 different MR-contrasts as channels, and the labelmap encodes 5 different label classes, i.e. $L(x) \in \{0, \ldots, 4\}$.

In the first step, an initial test-specific probabilitiy $p^c_{\mathrm{GMM}}(I)$ is created for each class c, by testing with a Gaussian Mixture Model of local multi-channel intensity for the class c, which is trained on all training data. These probabilities are then used to augment the original input data as additional channels. This can be seen as pre-processing for each image I, and we redefine $I = [I, p_{\mathrm{GMM}}(I)]$ to denote the resulting 9-channel 3D image for the following.

Based on the augmented input, we train randomized classification trees with context-aware features. Following the atlas forest scheme, each tree is trained only on an individual atlas A_i. A set of n such trees forms an atlas forest $a_i = \{T_i^k\}_{k=1:n}$. The training uses axis-aligned features and information gain as splitting criterion. Randomization is introduced via random sampling of the feature space by uniformly drawing feature types and parameters for the 3 randomized feature types: 1) Intensity difference between location of interest x in channel I_{j_1} and an offset point $x+v$ in channel I_{j_2}; 2) Difference between intensity means of a cuboid around x in I_{j_1} and a cuboid around $x+v$ in I_{j_2}; 3) Intensity range along a 3D line between x and $x+v$ in I_j. For cross-channel features 1) and 2), both I_{j_1} and I_{j_2} are drawn either from intensity or probability channels.

At test time, the image I is labeled by each atlas forest a_i, resulting in N candidate labelmaps $\hat{L}_i(x) = \arg\max_c \sum_j p_{T_i^j}(c|x, I)$. This is in contrast to the original atlas forests [6], which averages the probabilistic estimates of the AFs into a single prediction $\hat{L}(x) = \arg\max_c \sum_i \sum_j p_{T_i^j}(c|x, I)$.

2.2 Implicit Atlas Weighting by STAPLE

Given the set of candidate estimates $\hat{L}_i$, STAPLE [12] performs an Expectation-Maximization (EM) algorithm to estimate the conditional probability of the hidden true segmentation $p(L(x) = c|\{\hat{L}_i\}, \{\theta_i\})$, as well as the corresponding performances θ_i of the individual segmentations, modeled as confusion matrices. Starting from initial estimates for $\{\theta_i\}$, STAPLE iterates in standard EM manner until convergence, with the final segmentation estimate being $\hat{L}_\mathrm{S}(x) = \arg\max_c p(L(x)=c|\{\hat{L}_i\}, \{\theta_i\})$, with

$$p\left(L(x)=c \mid \{\hat{L}_i\}, \{\theta_i\}\right) = \frac{p\left(L(x)=c\right) \prod_i p\left(\hat{L}_i(x) \mid L(x)=c, \theta_i\right)}{\sum_{c'} p\left(L(x)=c'\right) \prod_i p\left(\hat{L}_i(x) \mid L(x)=c', \theta_i\right)} \quad (1)$$

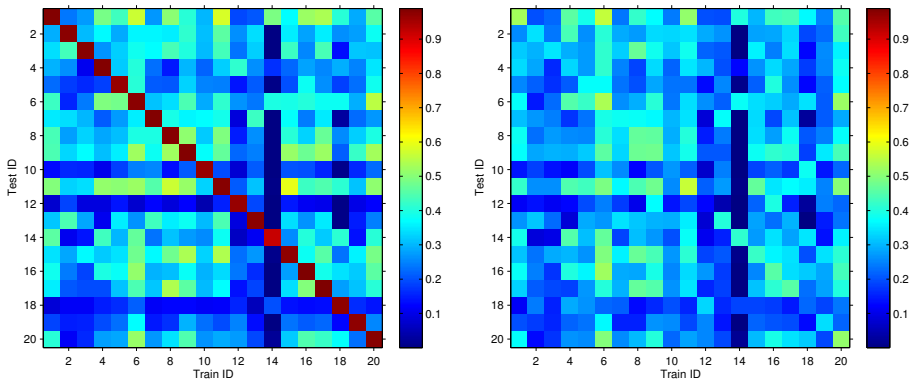
In the numerator, the prior $p(L(x)=c)$ is weighted by the probability of correct prediction of c by the candidate segmentation $\hat{L}_i(x)$, according to its estimated performance θ_i. Since the estimates $\hat{L}_i$ are directly associated to the atlases A_i via the atlas-forest classifiers a_i, this results in an implicit weighting of the training atlas images according to their estimated relevance. Please note that the performance for a_i is in general not the same for different classes.

3 Evaluation

After providing details about data and setup, we present two experiments: In Sec. 3.3 we evaluate the quality of STAPLE performance estimation, and in Sec. 3.4 we compare the results of a standard forest (Std. Forest) approach as in [13], AFs with probabilistic averaging as fusion (AF-PrAvg) as in [6], and the proposed AFs with atlas weighting by STAPLE (AF-STAPLE).

3.1 Data

The evaluation is performed on the real data from the NCI-MICCAI BraTS 2013 Challenge [4], which consists of 3 datasets: training, leaderboard and challenge. The training data, for which the reference manual segmentations are available, consists of 20 high-grade (HG) and 10 low-grade (LG) cases. Leaderboard has 21 HG and 4 LG cases, and challenge has 10 HG cases. We refer to the leaderboard and challenge data, for which the reference labelmaps are not known, as evaluation data. The actual labelmaps contain 5 classes, however, the challenge evaluation is performed on three "regions", which combine the classes to: *complete tumor*, *tumor core*, and *enhancing tumor*. For each case, 4 different MR contrasts are given as input data: contrast enhanced T1, T1, T2 and FLAIR. As additional pre-processing, we perform inhomogeneity correction by [14], set the median of each channel to a fixed value (1000), and downsample the images by factor of two with nearest-neighbor interpolation. Quantitative evaluation for all experiments is performed by submitting to the BraTS challenge system.



(a) Actual AF Performance (b) STAPLE Performance Estimate

Figure 1: Quality of STAPLE performance estimate on high-grade training data: (a) Actual Dice scores ($\mathbf{D}$) (average over all classes) resulting from testing on individual AFs (diagonal: testing on training images, note the over-training effect). (b) Dice scores *estimated* by STAPLE ($\mathbf{D}_{\mathrm{E}}$). Rows show performance of different AFs for a given test image, which is relevant for atlas weighting: e.g. AF-14 seldom performs well. The similarity between (a) and (b) shows that STAPLE performance estimates have high-quality: excluding the diagonal, average correlation of corresponding rows of $\mathbf{D}$ and $\mathbf{D}_{\mathrm{E}}$ is 0.87.

3.2 Implementation and Parameter Settings

For the basic tree training, we use the method as described in Section 2.1 with the same settings as in [13]. To replicate the results from [13] for Std. Forest, we train 60 trees per forest for HG and LG. For Std. Forest, we perform random subsampling of the background class with a sampling rate of 0.2. For Atlas Forests, we train 3 trees per AF, resulting in 60 trees for HG and 30 for LG. To perform STAPLE, we use the implementation from `http://www.crl.med.harvard.edu/software/STAPLE/index.php` with the default settings.

3.3 Quantifying the Quality of Performance Estimates by STAPLE

This experiment evaluates the ability of STAPLE to predict the performance of individual AFs for brain tumor data, with results summarized in Fig. 1. In this context, performance describes the accuracy of the prediction by an individual classifier. For this task, we use the HG training dataset, and train an AF a_i for each case. Then with each a_i we generate a set of segmentations $\{\hat{L}_i^k\}$ for each image I_k.

First, we measure the actual performance of the individual AFs in this scenario by computing the average Dice score per class for each $\hat{L}_i^k$ compared to the reference manual segmentation L_i (Fig. 1a). The rows of the matrix in Figure 1a

Method	Training All			Evaluation All		
	complete	core	enhancing	complete	core	enhancing
Std. Forest	75.0±15.2	63.8±29.1	44.9±36.8	74.3±16.4	62.7±28.0	51.6±32.5
AF Pr.Avg.	64.2±30.2	50.0±33.8	40.1±36.9	64.3±29.4	46.9±32.7	43.1±33.2
AF STAPLE	76.6±17.3	62.6±23.8	47.0±35.4	76.5±18.0	62.9±25.2	52.1±31.5

Table 1: Quantative summary of results on complete BraTS 2013 data, including *training* (leave-1-out validation), and *evaluation*. Please see also Fig. 2.

show the performance of different AFs for a given test image, which is relevant for atlas weighting. For example, one can observe that AF-14 seldom performs well. We observe a high variance of performance for the individual AFs, which was the initial motivation for this work. Please note the high values on the diagonal (training and testing on the same image), which shows the over-training property of AFs.

Second, we measure the ability of STAPLE to *estimate* the performance of the individual AFs, *without access to reference labels or any prior information* (Fig. 1b). For each test case k we apply STAPLE to candidate segmentations $\{\hat{L}_i^k\}_{i \neq k}$, which yields performance estimates for each AF a_i. These are quantified by computing (excluding the diagonal) the correlation of the actual and estimated performance matrix (0.91), and the average correlation of corresponding rows (0.87), showing the high accuracy of STAPLE for this task.

3.4 Evaluation on BraTS 2013

We use same settings for all experiments, but two different protocols for training and evaluation data. For training data, we perform a leave-1-out experiment to simulate a realistic scenario. For the evaluation data, we use all available training atlases. For each method, we separately train and test for HG and LG.

The results are summarized in Fig. 2 and Tab. 1, and seem consistent across the different data subsets. We observe that as expected, the original AF method with probabilistic averaging (AF-PrAvg) [6] has significantly reduced accuracy compared to the baseline (Std. Forest) which uses the same basic classifier. The proposed AF-STAPLE which performs atlas weighting has the desired effect of recovering the performance to the level of the original classifier, cf. Tab. 1, while keeping the computational efficiency advantages of the MALP framework.

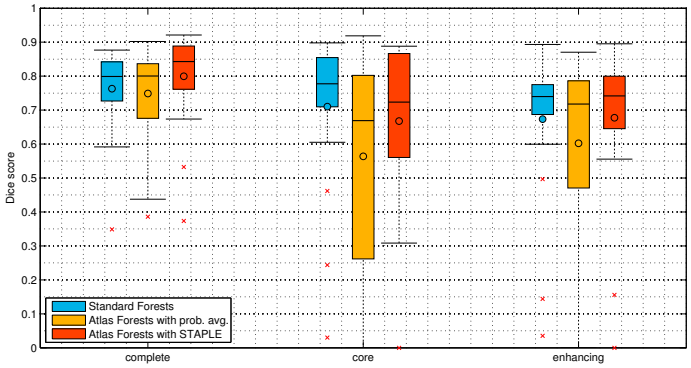
4 Discussion and Summary

We propose a segmentation method which retains the advantages of CB-MALP (increased efficiency, ability for atlas selection) but can be applied to general scenarios, such as brain tumor segmentation. The results show that even in such settings, in which other CB-MALP methods are shown to fail, the proposed approach is capable of the same accuracy as the standard learning scheme, while using the same basic classifier method. As future work, it would be interesting to

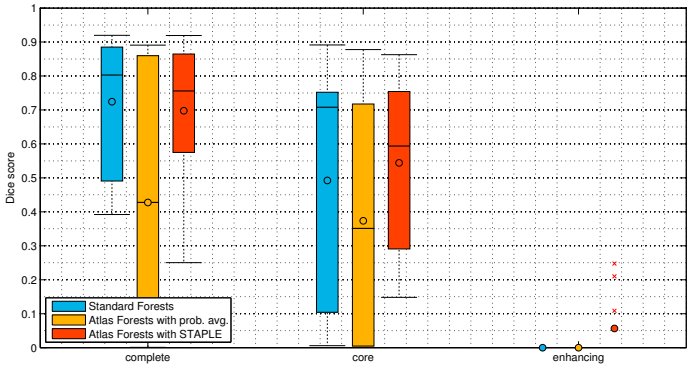
consider alternative classifiers, which are potentially more tuned towards specific problems, and evaluate the effect of the proposed framework. Also, different alternative weighting methods could be used, e.g. the probabilistic version of STAPLE from [7], which might be more suitable for classifier-based predictions. We believe that in the light of growing annotated databases, the ability to learn from more similar data has the potential to provide increased accuracy, especially for under-represented outlier cases.

References

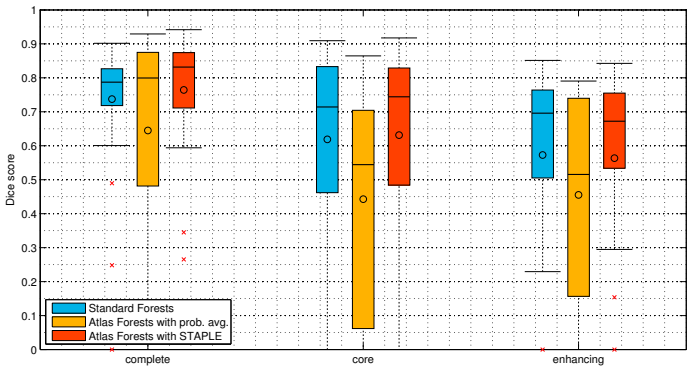
1. Landman, B., Warfield, S., eds.: MICCAI Workshop on Multi-Atlas Labeling. (2012)
2. Asman, A., Akhondi-Asl, A., Wang, H., Tustison, N., Avants, B., Warfield, S.K., Landman, B.: MICCAI 2013 Segmentation Algorithms, Theory and Applications (SATA) Challenge Results Summary. In: MICCAI Challenge Workshop on Segmentation: Algorithms, Theory and Applications (SATA). (2013)
3. Wolz, R., Chu, C., Misawa, K., Mori, K., Rueckert, D.: Multi-organ Abdominal CT Segmentation Using Hierarchically Weighted Subject-Specific Atlases. In: MICCAI. (2012)
4. Menze, B., et al.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). http://hal.inria.fr/hal-00935640 (2014)
5. 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 **46**(3) (2009) 726–738
6. Zikic, D., Glocker, B., Criminisi, A.: Atlas encoding by randomized forests for efficient label propagation. In: MICCAI. (2013)
7. Akhondi-Asl, A., Warfield, S.: Simultaneous truth and performance level estimation through fusion of probabilistic segmentations. IEEE TMI (2013)
8. 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
9. Heckemann, R., Hajnal, J., Aljabar, P., Rueckert, D., Hammers, A., et al.: Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage **33**(1) (2006) 115–126
10. Rousseau, F., Habas, P., Studholme, C.: A supervised patch-based approach for human brain labeling. IEEE TMI **30**(10) (2011) 1852–1862
11. Coupé, P., Manjón, J., Fonov, V., Pruessner, J., Robles, M., Collins, D.: Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. NeuroImage **54**(2) (2011) 940–954
12. Warfield, S., Zou, K., Wells, W.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE TMI **23**(7) (2004) 903–921
13. Zikic, D., Glocker, B., Konukoglu, E., Shotton, J., Criminisi, A., Ye, D., Demiralp, C., Thomas, O.M., Das, T., Jena, R., Price, S.J.: Context-sensitive classification forests for segmentation of brain tumor tissues. In: MICCAI 2012 Challenge on Multimodal Brain Tumor Segmentation (BraTS). (2012)
14. Tustison, N., Gee, J.: N4ITK: Nick's N3 ITK implementation for MRI bias field correction. The Insight Journal (2010)



(a) Training-HG: leave-1-out



(b) Training-LG: leave-1-out



(c) BraTS Evaluation

Figure 2: Evaluation on BraTS 2013 data: Leave-1-out experiment on (a) high-grade (HG) and (b) low-grade (LG) cases on the training data. (c) results on the evaluation data (*leaderboard* and *challenge*). Please see also Tab. 1.