Estimation of the Prior Distribution of Ground Truth in the STAPLE Algorithm: An Empirical Bayesian Approach

Alireza Akhondi-Asl, Simon K. Warfield

Computational Radiology Laboratory, Department of Radiology,Children's Hospital,
300 Longwood Avenue, Boston, MA, 02115, USA

Abstract. We present a new fusion algorithm for the segmentation and parcellation of magnetic resonance (MR) images of the brain. Our algorithm is a parametric empirical Bayesian extension of the STAPLE algorithm which uses the observations to accurately estimate the prior distribution of the hidden ground truth using an expectation maximization (EM) algorithm. We use IBSR dataset for the evaluation of our fusion algorithm. We segment 128 principle gray and white matter structures of the brain using our novel method and eight other state-of-the-art algorithms in the literature. Our prior distribution estimation strategy improves the accuracy of the fusion algorithm. It was shown that our new fusion algorithm has superior performance compared to the other state-of-the-art fusion methods in the literature.

1 Introduction

Fusion algorithms have been widely used in variety of medical image segmentation problems, in particular, in brain segmentation techniques. The key purpose of such algorithms is to match multiple templates to the target image and fuse their corresponding segmentations to have an estimate of the hidden ground truth. The simplest way to fuse the templates, known as majority voting, is to count the number of votes for each label and assign the label with the highest number of votes to the voxel. The key assumption in this method is that the votes are equally weighted, however, it is known that the templates or in the general case, raters, have different accuracies. Therefore, considering their performance variations may lead to a more accurate estimation of the hidden ground truth. To this end, two categories of fusion methods have been introduced in the literature, that is, intensity based weighted voting methods and STAPLE algorithm with its extensions and variations. In the former approach, intensity similarity of the target image and templates are used to estimate the performance of the raters, where intensity similarities are considered as the weights of the decision of each template at each voxel. Moreover, mean square error and normalized cross correlation based methods are mainly used as the similarity metric in this type of approach. However, these algorithms can be very sensitive to the intensity normalization and cannot compensate some of the intrinsic weaknesses of the

majority voting approach. [1–6].
In the latter method, the performance of the raters and the hidden ground truth are estimated iteratively using an Expectation-Maximization (EM) algorithm [7]. The method was first introduced for the estimation of the performance of the raters. Since then, it is being used in many applications as the fusion algorithm for the estimation of the hidden ground truth. There are many extensions of the algorithm, including SIMPLE, COLLATE, and STAPLER [8–10]. COLLATE algorithm is based on the STAPLE which considers the consensus level at each voxel for the estimation of the hidden ground truth. A similar concept was discussed by Rohlfing et al. and a simple solution was introduced, which is known as the STAPLE with assigned consensus region [11]. The idea discussed in [11] is to restrict the STAPLE algorithm to perform on the voxels where there is a disagreement between raters, otherwise, assign the corresponding label for the other voxels [11]. In this way, large consensus regions do not have any effect on the estimation of the performance parameters. STAPLER, another extension of the STAPLE, is designed to deal with the missing and also repeated segmentations. The algorithm uses training data to improve the estimation of the ground truth and performance parameters.
In the standard Bayesian analysis the prior distribution is assumed to be known and the observations do not re-scales the prior distribution. Therefore, in all of the STAPLE based algorithms, the prior on the ground truth needs to be set in the same manner. However, in practice this is usually discarded and the observations are used to estimate the prior distribution. While it is mentioned in the original STAPLE paper that the prior can be set using a probabilistic atlas, usually the decisions by the raters (observations) are used to estimate a global prior for each label.
In this paper we introduce a novel framework to estimate the prior using a parametric empirical Bayesian procedure. In our approach we estimate both performance parameters and hidden ground truth by updating the prior distribution of the ground truth using an EM algorithm. We show that this approach improves the accuracy of the fusion algorithm in the estimation of the hidden ground truth as well as the performance parameters.

2 Methods

Similar to any other fusion problem, we are interested in the estimation of $\mathbf{T}$, the hidden ground truth of the target image $\mathbf{I}$. It is assumed that T_i, the true label at voxel $i \in \{1, \ldots, i, \ldots, N\}$, is one of the labels $s \in \{1, \ldots, S\}$. In addition, we assume that J independent segmentation of the target image I are available where their performances are unknown. STAPLE is known to be one of the fusion methods that estimates the performance of the segmentations (raters) and utilize them in the estimation of the ground truth [7]. In the STAPLE algorithm, we are interested in the maximization of $f(\mathbf{D}, \mathbf{T}|\boldsymbol{\theta}, \boldsymbol{\rho})$, the probability density function of the complete data. In this formulation, $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_j, \ldots, \boldsymbol{\theta}_J\}$ indicates performance parameters where $\boldsymbol{\theta}_j$ is a matrix of size $S \times S$ and $\theta_{js's} =$

$f(D_{ij} = s'|T_i = s)$. In addition, $\mathbf{D}$ is the decision matrix of size $N \times J$ where D_{ij} is the decision of rater j at voxel i. Also, $\boldsymbol{\rho}$ is the prior matrix of size $S \times N$ where $\rho_{si} = f(T_i = s)$ is the prior probability of the label s at voxel i.
In this formulation, both $\mathbf{T}$ and $\boldsymbol{\theta}$ are unknown. Thus, to estimate the unknown parameters the EM algorithm is used. In [7] authors assume that the prior $\boldsymbol{\rho} = \hat{\boldsymbol{\rho}}$ is known. They have suggested different approaches to estimate the prior and they have used the global or spatially fixed prior. However, in our new approach and using the empirical Bayesian approach, we assume that $\boldsymbol{\rho}$ is another set of unknown parameters which controls the shape of the prior distribution of the hidden ground truth. To solve the problem, we use the EM algorithm to iteratively estimate the ground truth, the performance parameters, and the prior distribution of the hidden ground truth. Using voxel-wise independence assumption and given $\boldsymbol{\theta}^t$ and $\boldsymbol{\rho}^t$, the estimation of the performance parameters and the distribution of the prior at the step t, the function Q is maximized iteratively:

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\rho}|\boldsymbol{\theta}^t, \boldsymbol{\rho}^t) &= E\left[\log f(\mathbf{D}, \mathbf{T}|\boldsymbol{\theta}^t, \boldsymbol{\rho}^t)\right] \\ &= \sum_i \sum_{T_i} \log f(\mathbf{D}_i, T_i|\boldsymbol{\theta}, \boldsymbol{\rho}_i) f(T_i|\mathbf{D}_i, \boldsymbol{\theta}^t, \boldsymbol{\rho}_i^t) \end{aligned} \tag{1}$$

The weight variable, W_{si}^t is computed using the following equation:

$$\begin{aligned} W_{si}^t &= f(T_i = s|\mathbf{D}_i, \boldsymbol{\theta}^t, \boldsymbol{\rho}_i^t) \\ &= \frac{\left[\prod_j f(D_{ij}|T_i = s, \boldsymbol{\theta}_j^t)\right] f(T_i = s|\boldsymbol{\rho}_i^t)}{\sum_{s'} \left\{\left[\prod_j f(D_{ij}|T_i = s', \boldsymbol{\theta}_j^t)\right] f(T_i = s'|\boldsymbol{\rho}_i^t)\right\}} \\ &= \frac{\left[\prod_j \theta_{jD_{ij}s}^t\right] \rho_{si}^t}{\sum_{s'} \left\{\left[\prod_j \theta_{jD_{ij}s'}^t)\right] \rho_{s'i}^t\right\}} \end{aligned} \tag{2}$$

There are two sets of parameters that should be optimized. To this end, Eq. 1 is expanded as:

$$\begin{aligned} &Q(\boldsymbol{\theta}, \boldsymbol{\rho}|\boldsymbol{\theta}^t, \boldsymbol{\rho}^t) = \sum_i \sum_{T_i} W_{si}^t \log f(\mathbf{D}_i, T_i|\boldsymbol{\theta}, \boldsymbol{\rho}_i) = \\ &\sum_i \sum_s W_{si}^t \log\left[f(\mathbf{D}_i|T_i = s, \boldsymbol{\theta}) f(T_i = s|\boldsymbol{\rho}_i)\right] = \\ &\sum_i \sum_s W_{si}^t \log(\prod_j \boldsymbol{\theta}_{jD_{ij}s}) + \sum_i \sum_s W_{si}^t \log(\rho_{si}) \end{aligned} \tag{3}$$

The first part of Eq. 3 is related to the performance parameters and is optimized using the constraint $\sum_{s'} \theta_{js's} = 1$. Thus, maximization of 3 with respect to the performance parameters lead to:

$$\theta_{js's}^{t+1} = \frac{\sum_{Dij=s'} W_{si}}{\sum_i W_{si}} \tag{4}$$

For the prior distribution parameters, we have the constraint that $\sum_s \rho_{si} = 1$. Thus, using the Lagrange multiplier the following function is optimized with respect to the ρ_{si}:

$$Q'(\boldsymbol{\rho}_i, \lambda) = \sum_s W_{si}^t log(\rho_{si}) + \lambda \sum_s \rho_{si} \tag{5}$$

It can be seen that:

$$\frac{\partial Q'(\boldsymbol{\rho}_i, \lambda)}{\partial \rho_{si}} = \frac{W_{si}^t}{\rho_{si}} + \lambda = 0 \tag{6}$$

Using the constraint, it can be seen that $\lambda = -\sum_s W_{si} = -1$ which leads to the following estimation of the prior for the step $t+1$:

$$\rho_{si}^{t+1} = W_{si}^t \tag{7}$$

This means that the best estimation of the prior at each step is the estimated ground truth, given the performance and prior parameters at the previous step.

2.1 Relation to the Classic STAPLE

To see the relation to the classic STAPLE algorithm, we build a Maximum A Posteriori (MAP) formulation for the problem of estimation of the prior distribution. To this end, we consider a Beta distribution for each ρ_{si} with parameters α_{si} and β_{si} and reformulate Q' as follows:

$$Q'_{MAP}(\boldsymbol{\rho}_i, \lambda) = Q'(\boldsymbol{\rho}_i, \lambda) + \gamma \log \left(\prod_s {\rho_{si}}^{(\alpha_{si}-1)} (1-\rho_{si})^{(\beta_{si}-1)} \right) \tag{8}$$

It is known that the Beta function takes its maximum at $\frac{\alpha_{si}-1}{\alpha_{si}+\beta_{si}-2}$. If $\hat{\rho}_{si}$ is the prior probability of T_i at voxel i for the label s in the classic STAPLE algorithm, we set the parameters of the Beta function to take its maximum at this point. In the other words,

$$\frac{\alpha_{si}-1}{\beta_{si}-1} = \frac{\hat{\rho}_{si}}{1-\hat{\rho}_{si}} \tag{9}$$

Taking the derivative with respect to the ρ_{si} leads to:

$$\frac{\partial Q'_{MAP}(\boldsymbol{\rho}_i, \lambda)}{\partial \rho_{si}} = \frac{W_{si}^t}{\rho_{si}} + \lambda + \frac{\gamma(\alpha_{si}-1)}{\rho_{si}} - \frac{\gamma(\beta_{si}-1)}{1-\rho_{si}} = 0 \tag{10}$$

Solving the equation for λ, leads to the following equation:

$$\rho_{si} = \frac{W_{si}^t + \gamma(\alpha_{si}+\beta_{si}-2) + \gamma\frac{\beta_{si}-1}{\rho_{si}-1}}{1 + \sum_{n'} \gamma(\alpha_{n'i}+\beta_{n'i}-2) + \sum_{n'} \gamma\frac{\beta_{n'i}-1}{\rho_{n'i}-1}} \tag{11}$$

Using the relation between α_{si} and β_{si} it is easy to see that:

$$\rho_{si} = \frac{W_{si}^t + \gamma \frac{(\beta_{si}-1)(\rho_{si}-\hat{\rho}_{si})}{(1-\hat{\rho}_{si})(\rho_{si}-1)}}{1 + \sum_{n'} \gamma \frac{(\beta_{n'i}-1)(\rho_{n'i}-\hat{\rho}_{n'i})}{(1-\hat{\rho}_{n'i})(\rho_{n'i}-1)}} \tag{12}$$

In the extreme situation, if $\gamma \to \infty$ and $\beta_{si} = K$, where K is a constant, it can be seen that $\rho_{si} = \hat{\rho}_{si}$ is the solution of the Eq. 12 which is similar to the classic STAPLE algorithm. In the other words, classic STAPLE is the MAP solution of our new algorithm when the prior weight is very large.

3 Results

Table 1. Comparison of Dice coefficients for the proposed methods. Average segmentations of 128 brain structures for the IBSR datasets. M_1: our parametric empirical Bayesian approach, M_2: STAPLE with assigned consensus region [7], M_3: COLLATE [9], M_4: method proposed by Artaechevarria et al. [6], M_5: method proposed by Sabuncu et al. [5], M_6: Majority voting, M_7: SIMPLE [8], M_8: STAPLER [10], M_9: STAPLE without assigned consensus region [7].

	M1	M2	M3	M4	M5	M6	M7	M8	M9
1	**0.752**	0.743	0.731	0.711	0.704	0.697	0.690	0.685	0.638
2	0.697	0.697	0.678	0.687	0.685	0.678	0.674	0.622	0.580
3	**0.715**	0.707	0.697	0.685	0.688	0.679	0.666	0.662	0.622
4	**0.687**	0.679	0.675	0.652	0.656	0.644	0.651	0.642	0.605
5	**0.738**	0.732	0.719	0.709	0.707	0.702	0.695	0.669	0.623
6	**0.647**	0.645	0.638	0.606	0.600	0.598	0.590	0.609	0.576
7	0.726	**0.727**	0.708	0.695	0.694	0.686	0.669	0.670	0.632
8	0.691	0.691	0.680	0.673	0.663	0.654	0.651	0.641	0.603
9	**0.716**	0.713	0.701	0.698	0.699	0.688	0.683	0.644	0.600
10	**0.700**	0.693	0.688	0.686	0.676	0.668	0.655	0.640	0.600
11	0.696	**0.698**	0.678	0.661	0.661	0.652	0.642	0.654	0.616
12	**0.670**	0.669	0.657	0.644	0.638	0.627	0.621	0.640	0.606
13	**0.700**	0.699	0.698	0.674	0.676	0.663	0.664	0.640	0.601
14	**0.745**	0.739	0.727	0.716	0.717	0.707	0.702	0.664	0.615
15	**0.735**	0.732	0.734	0.703	0.709	0.693	0.687	0.665	0.618
16	**0.729**	0.721	0.713	0.701	0.698	0.687	0.689	0.661	0.618
17	**0.716**	0.714	0.716	0.682	0.684	0.668	0.675	0.657	0.614
18	0.724	0.724	0.712	0.683	0.678	0.675	0.672	0.666	0.625
Mean	**0.710**	0.707	0.697	0.681	0.680	0.670	0.665	0.652	0.611
STD	0.027	0.026	0.026	0.027	0.029	0.028	0.028	0.019	0.016
MAX	**0.752**	0.743	0.734	0.716	0.717	0.707	0.702	0.685	0.638
MIN	**0.647**	0.645	0.638	0.606	0.600	0.598	0.590	0.609	0.576
p-value		0.002	1e-06	1e-10	1e-9	1e-12	1e-12	1e-11	1e-13

(a) (b) (c)

Fig. 1. Illustration of IBSR Multi-Atlas Segmentation Results. Comparison of segmentation of pars triangularis (F3t) generated by (a) STAPLE with assigned consensus region as the closest algorithm to our method [7]. (b) our method: parametric empirical Bayesian approach. (c) expert manual segmentation in a coronal image of a representative IBSR dataset. It can be seen that our method is superior to the classic STAPLE algorithm.

We have evaluated our method and some of the state-of-the-art fusion algorithms for the automatic segmentation of Internet Brain Segmentation Repository (IBSR) MRI datasets. The database includes 18 T1-weighted volumetric images with slightly different voxel sizes and their corresponding manual segmentation and parcellation. The MR brain data sets and their manual segmentations were provided by the Center for Morphometric Analysis at Massachusetts General Hospital and are available at http://www.cma.mgh.harvard.edu/ibsr/. The volumetric images have been positionally normalized into the Talairach orientation (rotation only). In addition, bias field has been corrected for these data. There are two sets of manual segmentations for each one of the subjects: manual segmentation of 34 principle gray and white matter structures of the brain and parcellation results of 96 structures in Cerebral Cortex. Thus, for each subject combining these two manual segmentation sets leads to $34 - 2 + 96 = 128$ manually segmented structures.
We have used the leave-one-out strategy for the evaluation. In the other words, for each dataset, we have used the other 17 datasets (templates) for the segmentation of the 128 structures. To this end and in the first step, we registered all of the templates to the each one of the images (target images). Based on our experiments and within non-rigid-registration algorithms in the literature, SyN was chosen for the alignment of the templates to the target images [12]. For the next step, the output transformations of the non-rigid-registrations were applied to the corresponding label maps. The output of this step is fed in to the fusion algorithms. Finally, our novel method and some of the state-of-the-art fusion algorithms have been used for the segmentation of each one of the 128 structures. Methods that have been used for the comparison are: STAPLE with and without assigned consensus region, SIMPLE, STAPLER, COLLATE, methods of Sabuncu et al. and Artaechevarria et al [7, 6, 9, 10, 8, 5].
For the STAPLE with and without assigned consensus region, we have used

the implementations in http://crl.med.harvard.edu/software/STAPLE/. For the methods of SIMPLE, COLLATE, and STAPLER we have used implementations in http://www.nitrc.org/projects/masi-fusion/ with default settings. We have implemented methods of Sabuncu et al. and Artaechevarria et al. These two methods are sensitive to the intensity differences between the target image and atlases. To have a fair comparison we normalized intensities of each atlas using the histogram matching approach. In addition, we could not use some of the algorithms for the multi-category whole brain segmentation. Some of the algorithms perform well for the binary cases and implementations of some others were not applicable for the multi-category whole brain segmentation with 128 structures. To this end, we have used the fusion algorithm for each structure independently and compared the fusion results. Table 1 shows the comparison of nine methods using the average Dice coefficient of 128 strictures. It can be seen that our method outperforms the other fusion algorithms ($p < 0.002$). We also carried out a power analysis using a two-tailed paired t-test, for detecting an effect of the size we identified $(0.71 - 0.707)/0.0036 = 0.833)$ for a sample with $N = 18$, at the significance level of $\alpha = 5\%$. We found the power $(\beta) = 91\%$. Figure 1 shows the comparison of segmentation of a sample structure (pars triangularis (F3t)) using our method and the STAPLE with assigned consensus region in a coronal image of a representative IBSR dataset. It can be seen that using parametric empirical Bayesian framework to estimate the distribution of prior on the ground truth has substantial impact on the outcome of the fusion algorithm.

4 Conclusions

The prior has an important role in determining the local optimum to which the Expectation-Maximization algorithm converges. It has become conventional in usage of STAPLE to estimate the distribution of prior on the ground truth from the input data assuming all inputs are equally reliable. Obviously, when they are not all equally reliable, this can lead to poor performance. We have presented a new Expectation-Maximization algorithm to simultaneously estimate the ground truth and atlas performance from a set of segmentations by updating the prior distribution of the hidden ground truth using an empirical Bayesian approach. The introduced approach is effective when there is little information available from which to set the prior probability of the true labels. In our introduced method, the prior is updated iteratively which makes it less sensitive to the initialization of the prior distribution as compared to the classic STAPLE, where the prior distribution is fixed. 128 brain structures from IBSR dataset has been used for the evaluation of our algorithm and eight other state-of-the-art fusion methods in the literature. Our new algorithm is robust and has superior performance as compared to the other methods. Also, it was shown that the classic STAPLE algorithm is the MAP solution of our method when the weight of the prior is very large. It is possible to use the same idea in the STAPLE based algorithms such as COLLATE and local MAP STAPLE [13].

Acknowledgments

This investigation was supported in part by NIH grants R01 EB008015, R01 LM010033, R01 EB013248, and P30 HD018655 and by a research grant from the Boston Children's Hospital Translational Research Program.

References

1. Lötjonen, J., Wolz, R., Koikkalainen, J., Thurfjell, L., Waldemar, G., Soininen, H., Rueckert, D.: Fast and robust multi-atlas segmentation of brain magnetic resonance images. NeuroImage **49**(3) (2010) 2352–2365
2. van Rikxoort, E., Isgum, I., Arzhaeva, Y., Staring, M., Klein, S., Viergever, M., Pluim, J., van Ginneken, B.: Adaptive local multi-atlas segmentation: Application to the heart and the caudate nucleus. Medical Image Analysis **14**(1) (2010) 39–49
3. Yushkevich, P., Wang, H., Pluta, J., Das, S., Craige, C., Avants, B., Weiner, M., Mueller, S.: Nearly Automatic Segmentation of Hippocampal Subfields in In Vivo Focal T2-Weighted MRI. NeuroImage (2010)
4. Wang, H., Suh, J., Das, S., Pluta, J., Altinay, M., Yushkevich, P.: Regression-based label fusion for multi-atlas segmentation
5. Sabuncu, M., Yeo, B., Van Leemput, K., Fischl, B., Golland, P.: A generative model for image segmentation based on label fusion. Medical Imaging, IEEE Transactions on **29**(10) (2010) 1714–1729
6. Artaechevarria, X., Munoz-Barrutia, A.: Combination strategies in multi-atlas image segmentation: Application to brain MR data. Medical Imaging, IEEE Transactions on **28**(8) (2009) 1266–1277
7. Warfield, S., Zou, K., Wells, W.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. Medical Imaging, IEEE Transactions on **23**(7) (2004) 903–921
8. Langerak, T., van der Heide, U., Kotte, A., Viergever, M., van Vulpen, M., Pluim, J.: Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (simple). Medical Imaging, IEEE Transactions on **29**(12) (2010) 2000–2008
9. Asman, A., Landman, B.: Robust statistical label fusion through consensus level, labeler accuracy and truth estimation (collate). Medical Imaging, IEEE Transactions on (99) (2011) 1779–1794
10. Landman, B., Asman, A., Scoggins, A., Bogovic, J., Xing, F., Prince, J.: Robust statistical fusion of image labels. Medical Imaging, IEEE Transactions on **31**(2) (2012) 512–522
11. Rohlfing, T., Russakoff, D.B., Maurer, Jr., C.R.: Expectation maximization strategies for multi-atlas multi-label segmentation. In Taylor, C., Noble, J.A., eds.: Information Processing in Medical Imaging. Volume 2732 of Lecture Notes in Computer Science., Berlin/Heidelberg, Springer-Verlag (Jul. 2003) 210–221 18th International Conference, IPMI 2003, Ambleside, UK, July 2003.
12. Avants, B., Yushkevich, P., Pluta, J., Minkoff, D., Korczykowski, M., Detre, J., Gee, J.: The optimal template effect in hippocampus studies of diseased populations. NeuroImage **49**(3) (2010) 2457–2466
13. Commowick, O., Akhondi-Asl, A., Warfield, S.K.: Estimating a reference standard segmentation with spatially varying performance parameters: Local map staple. IEEE Transactions on Medical Imaging (2012)