Optimal MAP Parameters Estimation in STAPLE Using Local Intensity Similarity Information

Subrahmanyam Gorthi, Alireza Akhondi-Asl, and Simon K. Warfield

***Abstract*—In recent years, fusing segmentation results obtained based on multiple template images has become a standard practice in many medical imaging applications. Such multiple-templates-based methods are found to provide more reliable and accurate segmentations than the single-template-based methods. In this paper, we present a new approach for learning prior knowledge about the performance parameters of template images using the local intensity similarity information; we also propose a methodology to incorporate that prior knowledge through the estimation of the optimal MAP parameters. The proposed method is evaluated in the context of segmentation of structures in the brain magnetic resonance images by comparing our results with some of the state-of-the-art segmentation methods. These experiments have clearly demonstrated the advantages of learning and incorporating prior knowledge about the performance parameters using the proposed method.**

***Index Terms*—Atlas-based segmentation, brain, label fusion, maximum-a-posteriori (MAP) formulation, medical imaging, MRI, segmentation, Simultaneous Truth and Performance Level Estimation (STAPLE).**

I. Introduction

It has been shown in many recent works that the automated segmentations obtained based on multiple template images provide more accurate segmentations than the single-template-based methods [1]–[10]. Multiple-templates-based segmentation can be defined as the alignment of a set of reference images with the corresponding segmentations to the target image to be segmented and followed by the fusion of those aligned segmentations to estimate the reference standard segmentation.

Fusion methods can be broadly classified into three categories: 1) voting-based methods [4]–[6], 2) distance-based methods [7], [10], and 3) statistically driven methods [1]–[3], [8], [11]–[13]. Voting-based methods assign a weight to the decisions made by each template regarding the probable output label at each voxel in the target image and finally select a label that satisfies certain optimal criteria. Distance-based methods compute the signed Euclidean distances to the contours of the structures, weigh those distances based on the similarity information, and finally assign a label that results in the least cumulative weighted Euclidean distance. On the other hand, the third category of statistical fusion methods simultaneously estimates both the probable output segmentation and performance parameters for each template, using an iterative approach.

Simultaneous Truth and Performance Level Estimation (STAPLE) is a widely used algorithm [1] that belongs to the third category of statistical fusion methods. The STAPLE algorithm not only generates the output segmentations (or reference standard), but also simultaneously rates the performances of the input segmentations. In practice, there are especially two specific scenarios where the STAPLE algorithm is widely used. First, it is used in order to generate the ground truth segmentations (also known as "reference standard") from multiple manual delineations prepared by multiple experts (or even the multiple delineations prepared the same expert at different times). Second, STAPLE algorithm is used for merging multiple automated segmentations that are obtained by registering multiple template images to a new target image and thereby generate more accurate segmentations for the target image than those individual segmentations from each template. It is this second scenario that we focus in the current manuscript.

The Expectation-Maximization (EM) approach used with the classical STAPLE algorithm guarantees convergence to a local optimum solution. However, if we can incorporate appropriate prior knowledge about the performance parameters of the templates into the maximum-a-posteriori (MAP) formulation of the STAPLE [12], [13], then it can provide more accurate estimations of both the reference standard and performance parameters.

MAP-based formulation of the STAPLE algorithm is used previously, for a different purpose, in order to merge the manual delineations made by multiple experts [12], [13]; it is used in the context of performing fusion with missing manual delineations for some of the structures of interest, in one or more template image. Such situation arises when some of the experts did not delineate all the structure of interest, but delineated only a subset of all the labels. To address such scenario, the authors in [12] and [13] proposed to incorporate this "missing" information into the STAPLE by appropriately constraining the performance parameters through the MAP formulation. As that approach is specifically designed to deal with the fusion problem in the presence of missing data, it does not distinguish between the performances of the regular templates without any missing data. The current manuscript addresses a completely different problem of learning prior knowledge about the performance parameters of automated segmentations obtained from multiple template images into the MAP formulation of the STAPLE algorithm.

Manuscript received December 31, 2014; revised March 12, 2015; accepted April 28, 2015. Date of publication April 30, 2015; date of current version September 1, 2015. This research was supported in part by NIH grants R01 EB013248 and R01 NS079788. The work of S. Gorthi was supported by the Swiss National Science Foundation under Grant P2ELP2_148892.

The authors are with the Computational Radiology Laboratory, Boston Children's Hospital, Harvard Medical School, Boston, MA 02115 USA (e-mail: Subrahmanyam.Gorthi@childrens.harvard.edu; Alireza.Akhondi-Asl@childrens.harvard.edu; simon.warfield@childrens.harvard.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JBHI.2015.2428279

In this paper, we introduce a general and powerful framework for learning prior knowledge about the performance parameters of each label in each template and for using that information to optimally set the MAP parameters of the STAPLE algorithm. More specifically, we propose here a new approach for learning the relationships between the local intensity similarities and the performance parameters of each label. Some of the previous works learn prior knowledge about performance parameters based on the training data [14] and also from the labels of the template images [13]. To the best of our knowledge, this is the first work that deals with learning prior knowledge about the performance parameters *from the intensity information of the template images*.

The method that we propose in this paper is an extension of the preliminary ideas that we have presented in a recent Workshop [15]. There are, however, substantial new contributions and extensions in the current manuscript compared to [15], and they are as follows. First, we modified the way we learn the relationships between the performance parameters and the similarity information. Second, unlike in [15], we compute the relationships locally, but not globally. Finally, we present here a comprehensive evaluation on ten subcortical structures in the brain MR images, and we also compare our results with many state-of-the-art fusion methods.

The remainder of this paper is organized as follows. Section II describes our new method and the proposed optimal MAP parameters estimation procedure. Section III presents a detailed evaluation of the proposed method for the segmentation of subcortical structures in brain MR images and comparisons with state-of-the-art methods. Finally, discussion and conclusions are presented in Section IV.

II. Methods

A. Regular EM-Based Formulation of STAPLE

As mentioned in the preceding section, the STAPLE algorithm takes multiple segmentation results obtained from multiple template images as the input; it then estimates both the final output segmentations and the performance parameters for each template image.

Let $D = \{D_1, \ldots, D_i, \ldots, D_N\}$ be a matrix of size $J \times N$, where J and N are, respectively, the number of templates and the number of voxels. In this matrix, $D_i = [D_{i1}, \ldots, D_{ij}, \ldots, D_{iJ}]'$ and D_{ij} is the label of the template j at voxel i. The goal here is to estimate the output segmentation $T = \{T_1, \ldots, T_i, \ldots, T_N\}$ and the performance parameters $\theta = \{\theta_1, \ldots, \theta_j, \ldots, \theta_J\}$ where θ_j is the matrix of size $S \times S$, $\theta_{js's} = f(D_{ij} = s' | T_i = s)$, and S is the number of segmentation labels.

Since both the output segmentations (T) and the performance parameters (θ) are unknown, the following complete data log-likelihood function is maximized iteratively using an EM algorithm:

$$Q(\theta|\theta^t) = \sum_i \sum_j \sum_s W_{si}^t \log(\theta_{jD_{ij}s}) \tag{1}$$

where W_{si}^t is the posterior probability of the reference standard segmentation T_i for label s.

The EM algorithm approaches the problem of maximizing the above log-likelihood function by proceeding iteratively with estimation and maximization steps. In the *estimation step*, the evaluation of $Q(\theta|\theta^t)$ requires the computation of posterior probability of T for each label s, and it is given by

$$\begin{aligned} P(T = s|D, \theta^t) &= \prod_i W_{si}^t \\ &= \prod_i \frac{P(T_i = s) \prod_j \theta_{jD_{ij}s}^t}{\sum_{s'} P(T = s') \prod_j \theta_{jD_{ij}s'}^t}. \end{aligned} \tag{2}$$

Given the estimated weight variables W_{si}^t, the new performance parameter θ^{t+1} at iteration number: $(t+1)$ are computed by *maximizing* the complete log likelihood function $Q(\theta|\theta^t)$.

The above EM formulation of the STAPLE algorithm guarantees convergence to a local optimum. However, incorporating appropriate prior knowledge about the performance parameters of the template images through the MAP formulation of the STAPLE algorithm could not only result in convergence to a global optimum (or strong local optimum), but also could result in more accurate estimation of both the performance parameters and the output segmentations. The following subsection presents beta distribution based MAP formulation of the STAPLE.

B. Beta Distribution Based MAP Formulation of STAPLE

The MAP formulation of the STAPLE algorithm can be expressed as

$$Q_{\text{MAP}}(\theta|\theta^t) = Q(\theta|\theta^t) + \gamma \log(p(\theta)) \tag{3}$$

where $p(\theta)$ is the prior probability of the performance parameters, and γ is the weighting parameter between the data term and of the MAP prior. As the performance parameters for each template and each label can be considered to be independent of each other [13], $p(\theta)$ can be expressed as a product of the probabilities of each performance parameter denoted by $p(\theta_{js's})$.

Similar to [13], in this paper, we use beta distribution $B_{\alpha,\beta}(x) = \frac{1}{Z} x^{\alpha-1}(1-x)^{\beta-1}$ for modeling the prior probabilities of each performance parameter. The main advantage of using beta distribution is that it facilitates modeling a variety of differently shaped performance characteristics by simply varying the two shape parameters: α and β; moreover, it is straightforward with the beta distribution to obtain its logarithm and derivatives that are required during the optimization procedure. Using beta distribution for modeling the prior probabilities of the performance parameters leads to the following expected value of the complete log-likelihood function:

$$\begin{aligned} Q_{\text{MAP}}(\theta|\theta^t) = &\sum_i \sum_j \sum_s W_{si}^t \log(\theta_{jD_{ij}s}) \\ &+ \gamma \sum_j \sum_{s'} \sum_s \Big[(\alpha_{js's} - 1) \log(\theta_{js's}) \\ &+ (\beta_{js's} - 1)(\log(1 - \theta_{js's})) \Big]. \end{aligned} \tag{4}$$

Notice that the computation of posterior probabilities depends only on the current estimates of θ^t, but not on the prior on these parameters; hence, the computation procedure for the posterior probabilities of the output segmentation T for each label s is same for both the EM-based and the MAP-based formulations of the STAPLE algorithm; the posterior probabilities are already presented in (2).

θ values that optimize (4) can be obtained by equating the derivatives of Q_{MAP} to 0 for each template image j; this results in the following system of equations:

$$\theta^t_{js's} = \frac{\gamma A_{s's} + \sum_{i:D_{ij}=s'} W^t_{si}}{\sum_{n'}\left(\gamma A_{n's} + \sum_{i:D_{ij}=n'} W^t_{si}\right)} \tag{5}$$

where

$$A_{n's} = \alpha_{jn's} + \beta_{jn's} + \frac{\beta_{jn's} - 1}{\theta_{jn's} - 1} - 2. \tag{6}$$

The above system of equations always has a unique solution and is known as *fixed point*. The solution scheme consists of an iterative process and is described in detail in [12].

In case of a binary segmentation problem (i.e., $s \in \{0, 1\}$), several simplifications can be made to the above system of equations, and it finally results in the following analytical closed-form solution [13]:

$$\begin{aligned} \theta^t_{jss} &= \frac{\sum_{i:D_{ij}=s} W^t_{si} + \gamma\left(\alpha_{jss} - 1\right)}{\sum_i W^t_{si} + \gamma\left(\alpha_{jss} + \beta_{jss} - 2\right)} \\ \theta^t_{j01} &= \left(1 - \theta^t_{j11}\right) \\ \theta^t_{j10} &= \left(1 - \theta^t_{j00}\right). \end{aligned} \tag{7}$$

In [12] and [13], the authors used the MAP solution for the specific problem of missing data. To this end, they used a set of empirically fixed parameters for all of the templates containing labels, to have priors with probability close to one for diagonal performance parameters, and close to zero for off-diagonal performance parameters. However, in this paper, we are interested in incorporating the prior knowledge about the performance parameters of each label in each template. The following section presents our proposed approach for achieving this goal, which is based on learning the relationships between the performance parameters and the image similarity information.

C. Learning Performance Parameters Versus Image Similarity Relations

In this paper, we consider the binary segmentation problem. Notice that, in case of binary segmentation, the diagonal elements of the performance matrix θ represent specificity and sensitivity [13], while the off-diagonal elements are (1-sensitivity) and (1-specificity); thus, we only need to learn prior knowledge about sensitivity and specificity. Please note that, in the remainder of this paper, when we say "performance parameters," we are actually referring to only the diagonal elements of the matrix θ (i.e., specificity and sensitivity).

A common underlying assumption for many fusion methods [4]–[7] is that the accuracy of segmentations obtained from a given template is proportional to its intensity similarity to the target intensity image. Similarly, we make here an assumption that if the intensity similarity of a template to the target intensity image is low, there is a high probability that its performance parameters are poor. This assumption is based on the observation that, a low intensity similarity can be an indication of significant anatomical differences between the template and the target intensity images, or (and) an indication of considerable error in registering the template to the target intensity image; since both of these scenarios could eventually reduce the accuracy of segmentation results obtained based on that particular template, we make the aforementioned assumption.

We then proceed further by learning the relationships between the performance parameters and the intensity information, by using all templates as our training data. The training procedure that we proposed in [15] for learning the prior knowledge is briefly as follows.

1) Select an image from the template database and treat it as the target image to be segmented (i.e., *pseudotarget image*). The rest of the images in the database are used as templates for that pseudotarget image.
2) Compute the *nonconsensus mask* for the pseudotarget image that contains only those voxels for which at least two template images disagree regarding output label and compute both the performance parameters over this mask.
3) Compute intensity similarities over the nonconsensus mask.
4) Repeat steps 1 to 3 for each image in the template database using a leave-one-out approach.
5) By the completion of step-4, for a database of J templates, we will have $J(J-1)$ pairs of sensitivity (or specificity) versus similarity values. Perform a robust linear regression analysis and obtain the final parameters representing the overall relation between the sensitivity (or specificity) and the image similarity.

In this paper, we propose the following modifications to the aforementioned learning approach:

1) Instead of learning the relationships over the entire image, we propose to learn them *locally*. This is based on the well-known observation that the intensity similarity between two images could vary significantly across different spatial locations, and thus, making inferences based on the local intensity similarity could result in more accurate results than the global intensity similarity.
2) In order to avoid introducing any undesired bias while estimating the relationships, unlike in the aforementioned approach, we do not use any mask; instead, we compute the similarity metric at each voxel, based on the intensity information at all the neighboring voxels that are present within the predefined radius (r_s) around that voxel.
3) Notice that, learning the relationships locally using the previously proposed approach in [15] requires performing robust linear regression at each voxel in the image; but, such approach becomes computationally very demanding with the increasing number of template images and image sizes. Hence, in this paper, we propose a new approach that estimates the MAP parameters directly based on the

similarity metric values, without requiring any regression analysis at each voxel.

The MAP parameters estimation procedure that we propose in this paper is presented in the following section.

D. MAP Parameters Estimation

As described in the preceding section, if the similarity between a template and the target intensity image is low, there is a high probability that the performance parameters of that template are low; similarly, we could expect high values of performance parameters (i.e., sensitivity and specificity) for the segmentations obtained from a template that has high intensity similarity to the target image.

The intensity similarity between two images can be estimated using various metrics like "Mean Square Error" and "Normalized Cross Correlation" (NCC). In this study, we use NCC as the intensity similarity metric; however, it is easy to notice that the proposed approach can be easily adapted to other similarity metrics as well.

Let φ_i^j represent the NCC value between the jth aligned template image and the target image, computed over a neighborhood patch of radius r_p that is centered at the ith voxel. It is known that φ_i^j varies between -1 and $+1$, whereas the values of the performance parameters vary between 0 and 1. In order to map high intensity similarity values to high performance parameters during the initialization, and also to map the range of NCC values to the range of performance parameter values, we first apply the following exponential-based transformation:

$$m_i^j = \frac{1}{1 + \mathrm{e}^{-A\left(\varphi_i^j - b\right)}} \tag{8}$$

where A and b are, respectively, the scale and the shift parameters that can be optimized for each specific problem. Notice that the above function, before applying the exponential-based transform, shifts φ_i^j by a value b, and then scales it by a factor A; thus, intuitively, this function not only maps NCC values from $[-1\ \ 1]$ to $(0\ \ 1]$, but also reduces the weight (or importance) given to φ_i^j values that are below b, and then scales the resultant values by a factor A.

We now present how we use m_i^j value for computing α_{jss} and β_{jss} parameters of the beta distribution, at each voxel, for each template j, and label s.

Notice that the mode of a beta distribution $B_{\alpha,\beta}(x)$ represents the x value where the distribution reaches a maximum value. In other words, the mode value can be interpreted as the "best guess" of what we are likely to see on a single realization of the target activity. Subjective estimates of the mode value are not only easier to elicit, but also more reliable than subjective estimates of the other characteristics (or parameters) such as mean, α and β values.

In our previous work [15], we assumed that a linear relationship exists between the mode values of the beta distribution of each performance parameter, and the intensity similarity metric. In the current study, we modify this somewhat strong assumption of "linear relationship" between the mode and NCC values by assuming a more general relationship presented in (8). The scale and the shift parameters of the exponential-based transformation in (8) gives more freedom in choosing the exact form (or shape) of the relationship between the mode values of the performance parameters, and the NCC values. In our current experiments, although we have used the same scale and shift values for all the voxels, the proposed framework facilitates optimizing these values individually for each label. To summarize, we assume here that the mode value of the beta distribution for the template j and voxel i occurs at m_i^j.

Furthermore, notice that the variance of the beta distribution indicates our confidence on the prior knowledge about the performance parameters that we learn based on the intensity similarities; in other words, small variance value of the beta distribution indicates high confidence on the prior knowledge about the performance parameters, and conversely, high variance value indicates less confidence on the prior knowledge. In all our experiments presented in this paper, we have empirically set the variance of the beta distribution to a fixed value ($1e-4$).

This implies that for each beta distribution, we know the mode and variance values, and the goal now is to obtain their equivalent α and β values as parameterized in (4). For this purpose, we use the method that was proposed in [16], and we now briefly summarize the derivation procedure.

Let m and σ^2, respectively, represent the mode and variance of the beta distribution. Since the mode occurs when the beta distribution reaches maximum, i.e., when the derivative is zero, the mode of the beta distribution parametrized in terms of α and β parameters is given by

$$m = \frac{\alpha - 1}{\alpha + \beta - 2}. \tag{9}$$

Similarly, the variance of the beta distribution is given by

$$\sigma^2 = \frac{\alpha\,\beta}{(\alpha+\beta)^2\ \ (\alpha+\beta+1)}. \tag{10}$$

Our goal now is to obtain the α and β values that result in the mode and variance values given by (9) and (10), respectively. This can be achieved through the standard rewriting and solving of the above two equations. A more detailed description of the relevant procedure can be found in [16].

For the convenience of notation, let us define an intermediate variable τ as

$$\tau = \frac{\sigma^2}{(1-m)^2}. \tag{11}$$

Then, the parameter β of the beta-distribution corresponds to the largest positive real root of the following cubic equation:

$$c_3\beta^3 + c_2\beta^2 + c_1\beta + c_0 = 0 \tag{12}$$

whose coefficients are given by

$$\begin{aligned} c_0 &= -12\tau m^3 + 20\tau m^2 - 11\tau m + 2\tau \\ c_1 &= 16\tau m^2 + (2 - 18\tau)m + 5\tau - 1 \\ c_2 &= -(7\tau + 1)m + 4\tau \\ c_3 &= \tau. \end{aligned}$$

The other shape parameter α of the fitted beta distribution is given by

$$\alpha = \frac{(\beta - 2)m + 1}{1 - m}. \tag{13}$$

To summarize, prior knowledge about the performance parameters of each template at each voxel is inferred based on the intensity information; this prior knowledge is incorporated into the MAP formulation of the STAPLE presented in (4), through α and β parameters of the distribution that were computed using (13) and (12), respectively.

Regarding the weighting parameter γ in (4), in all our experiments, we set its value to the average number of voxels present in the output label that is obtained based on the simple EM-based STAPLE algorithm; by this way, the two terms in the MAP formulation of the STAPLE will have approximately similar weight. While applying the MAP-STAPLE algorithm locally, for each voxel, over a neighborhood radius of r_s, we scale the γ value accordingly, using the empirically driven expression presented in [17], and thus, the new weight γ' is given by

$$\gamma' = \gamma \, \frac{N_w \ln(J)}{N} \tag{14}$$

where N is the total number of voxels in the image, J is the number of templates, and N_w is the number of voxels present within the cube of radius r_s.

Finally, we want to summarize the complete algorithm that we have described so far throughout this section.

1) Compute the local intensity similarity metric (NCC) at each voxel in the target image, for each aligned template image.
2) Compute the mode values corresponding to each similarity metric value (computed in step 1) using the exponential-based transformation presented in (8).
3) Compute the α and β parameters of the beta distributions corresponding to each mode value (computed in step 2) using the system of equations presented in (12) and (13).
4) Compute the weighting parameter γ' for a given weight (γ) and r_s using the expression presented in (14).
5) Solve iteratively the system of equations presented in (2) and (5).

Please note that, unlike in [15], we compute the intensity similarity metric (mentioned step-1) at a given voxel based on the intensity information at *all* the neighboring voxels present within the radius of r_s; on the other hand, like most of the STAPLE-based algorithms, we estimate the ground truth (mentioned in step-5) *only* at the nonconsensus voxels.

III. Experiments

In this section, we validate our new method in the context of segmentation of structures in the 3-D brain MR images. In addition, we compare the results from our proposed approach with the results from some of the state-of-the-art fusion methods.

A. Dataset

We utilize the IBSR brain dataset[1] of 18 healthy subjects for our experiments. It is a publicly available dataset that contains T1 intensity images of subjects, and the corresponding ground truth segmentations for various structures in the brain. We considered ten subcortical structures for our evaluation: 1) Left Thalamus, 2) Right Thalamus, 3) Left Caudate, 4) Right Caudate, 5) Left Putamen, 6) Right Putamen, 7) Left Pallidum, 8) Right Pallidum, 9) Left Hippocampus, and 10) Right Hippocampus.

B. Registration Procedure

The registration procedure that we followed in this paper is very similar to [19]. We started with a linear registration step for initial alignment; we linearly registered all the 18 brain images to a common template using FMRIB Software Library's (FSL) FLIRT with the following settings: nine-parameter, correlation ratio, trilinear interpolation; the common template was the "nonlinear MNI152," the nonlinear average template in MNI space used by FSL.

We then rigidly registered each of the 18 brain images in the MNI space to the rest of the 17 images in a leave-one-out manner, again using FLIRT with the following settings: six-parameter, correlation ratio, trilinear interpolation.

As a final registration step, we performed nonrigid registration between each pair of rigidly registered images, using the diffeomorphic demons registration algorithm proposed in [20]. For this purpose, we used the publicly available ITK implementation of the diffeomorphic demons registration [21]. As in [19], we used the following settings for this nonrigid registration: three multiscale-pyramid levels with iterations of 30, 20, and 10, respectively, smoothing sigma of 2.0 for the deformation field, and use of histogram matching prior to registration.

C. Fusion Methods and Parameters

We compare the results from our new method with the results obtained from various categories of existing fusion methods, namely, simple voting method, voting-based method that uses local intensity information ([4]), STAPLE-based methods that do not use any intensity information ([1], [13], [18]), and STAPLE-based method that uses local intensity information ([8]).

More specifically, we evaluate the segmentation results obtained from the following fusion methods:

1) MV
2) STAPLE [1]
3) COLLATE [18]
4) LWV [4]
5) LOP-STAPLE [8]
6) Empirical local MAP-STAPLE [13]
7) Our new optimal local MAP-STAPLE

[1]The MR brain datasets 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/.

TABLE I
PARAMETERS USED FOR LOCAL WEIGHTED VOTING (LWV) [4], LOGARITHMIC OPINION POOL-BASED STAPLE (LOP-STAPLE) [8], EMPIRICAL LOCAL MAP-STAPLE METHOD [13], AND OUR NEW OPTIMAL LOCAL MAP-STAPLE ARE DESCRIBED IN THIS TABLE.

Fusion Method	Parameters values	Description
LWV	$r_p = 3$	Half window size for computing intensity similarity
LOP-STAPLE	$r_p = 2$	Half window size for computing intensity similarity
	$A = 5$	Scale parameter
	$b = 0.8$	Shift parameter
Empirical Local MAP-STAPLE	$\alpha_{js's}\ (s' \neq s) = 1.5$	α value for nondiagonal elements of the Beta distribution
	$\alpha_{jss} = 5$	α value for diagonal elements of the Beta distribution
	$\beta_{js's}\ (s' \neq s) = 1.5$	β value for nondiagonal elements of the Beta distribution
	$\beta_{jss} = 5$	β value for diagonal elements of the Beta distribution
	$r_s = 7$	Half window size for computing performance parameters
Optimal Local MAP-STAPLE	$r_p = 4$	Half window size for computing intensity similarity
	$A = 3$	Scale parameter
	$b = 0.8$	Shift parameter
	$r_s = 7$	Half window size for computing performance parameters

For STAPLE [1] and COLLATE [18] methods, we used their default parameters presented in those respective papers. For empirical local MAP-STAPLE, we used the same parameters that were used in [13] for the templates that did not have any missing labels, i.e., we set the α and β values to 5 and 1.5, respectively, for the diagonal elements of the performance matrix for all the template images; notice that those particular values of α and β are equivalent to setting a mode value of 0.89, and variance of 0.02 for all the template images. We use same value of r_s (i.e., half window size for computing performance parameters) for both the empirical local MAP-STAPLE, and our optimal local MAP STAPLE, so that the comparison between these two methods will be fair.

Unlike the above three methods, LWV, LOP-STAPLE, and optimal local MAP-STAPLE have certain parameters, especially related to the intensity information, that need to be optimized. To this end, we optimized those parameters for each fusion method independently, by evaluating the Dice similarity coefficient of all ten structures, and for all 18 images in the dataset; for each fusion method, we have finally selected those parameters that resulted maximum overall Dice similarity coefficient. The parameters used for different fusion methods are described in Table I.

D. Evaluation Results

In this section, we evaluate our new "optimal local MAP-STAPLE" fusion method, by comparing it with the existing MV, STAPLE, COLLATE, LWV, LOP-STAPLE, and empirical local MAP-STAPLE methods. We perform the evaluation in the context of segmenting ten subcortical structures in the IBSR brain dataset of 18 images. We use the leave-one-out approach for template-fusion, i.e., for each target image, we combine the segmentation results obtained from the remaining 17 template images that are registered to the current target image.

As we have considered only the binary segmentation problem, all the ten structures are segmented independently. In order to speed up the fusion process, we computed the regions of interest for each structure, based on the labeled images of all templates, and then the images are cropped accordingly.

We use the average Dice similarity coefficient for comparison of the fusion methods. Furthermore, in order to evaluate the statistical significance of the results, we also perform two-sided Wilcoxon signed rank test (with significance level of 0.05) between each existing fusion method and the new method. Table II presents the average (mean) and standard deviation values of Dice similarity coefficients obtained from all fusion methods, for all the structures, along with the statistical tests results. Finally, Fig. 1 shows a representative segmentation obtained for one of the images, using MV, STAPLE, and the proposed method.

Based on the average Dice similarity coefficient values presented in Table II, the following observations can be made. The proposed optimal local MAP-STAPLE method provided the best overall segmentation results among all the seven fusion methods, resulting in an average Dice similarity coefficient of 82.66%. The best overall and structure-wise Dice similarity values are marked in bold in the table for an easy reference. When we look at the segmentation results structure-wise, the proposed method provided the best segmentation results for nine out of ten structures; for the other remaining structure (i.e., left hippocampus), segmentation results from LOP-STAPLE are slightly better than our new method. To summarize, optimal local MAP-STAPLE provided the best overall segmentation results, and it is followed by LOP-STAPLE, COLLATE, STAPLE, LWV, empirical local MAP-STAPLE, and MV, respectively.

Regarding the computational aspects, all the experiments are run on a 64-bit 10-core workstation with Intel Xeon 2.40-GHz processor, and 47-GB RAM. All the fusion methods, except COLLATE, are implemented in C++, with parallel processing. For COLLATE, we used the MATLAB-based implementation provided by the authors of [18]. Table II presents the average computational times per structure for each fusion method; notice that it took less than a minute (53 s) for the proposed algorithm. Thus, with the parallel implementation run on ten cores, when compared to the empirical local MAP STAPLE (48 s), there is only an additional overhead of around 5 s per structure for the proposed method.

In addition to the average Dice similarity results, Table II also presents various statistical metrics obtained based on two-sided

TABLE II
AVERAGE DICE SIMILARITY RESULTS, COMPUTATIONAL TIMES, AND STATISTICAL RESULTS FOR THE SEGMENTATION OF TEN SUBCORTICAL STRUCTURES, IN A DATASET OF 18 SUBJECTS.

Structure Name	MV	STAPLE	COLLATE	LWV	LOP-STAPLE	Empirical Local MAP-STAPLE	Optimal Local MAP-STAPLE
1. Left Thalamus	88.04%	88.25%	88.25%	88.12%	88.41%	87.98%	**88.45%**
2. Right Thalamus	87.25%	87.47%	87.55%	87.43%	87.74%	87.22%	**88.01%**
3. Left Caudate	83.02%	83.28%	83.17%	83.08%	83.33%	82.87%	**83.52%**
4. Right Caudate	82.03%	82.23%	82.12%	82.16%	82.31%	81.79%	**82.76%**
5. Left Putamen	85.49%	85.77%	85.93%	85.64%	85.98%	85.61%	**86.23%**
6. Right Putamen	84.77%	85.17%	85.25%	85.12%	85.57%	84.89%	**86.10%**
7. Left Pallidum	73.19%	75.06%	74.78%	74.48%	75.52%	74.49%	**76.74%**
8. Right Pallidum	70.64%	73.26%	73.90%	72.52%	74.27%	71.90%	**76.13%**
9. Left Hippocampus	77.66%	78.62%	78.80%	78.31%	**79.18%**	77.90%	79.17%
10. Right Hippocampus	77.79%	78.87%	79.04%	78.52%	79.40%	78.35%	**79.45%**
Average	80.99%	81.80%	81.88%	81.54%	82.17%	81.30%	**82.66%**
Standard Deviation	5.94%	5.17%	5.10%	5.39%	4.95%	5.47%	**4.57%**
Average computational time per structure	< 1 s	< 1 s	6.5 s	3.5 s	8.1 s	48.3 s	52.5 s
$p <$	1e-5	1e-5	1e-5	1e-5	0.0005	1e-5	
V	15927	12640	11574.5	15828	10563	15174	
C_H	0.014	0.008	0.008	0.009	0.005	0.012	
C_L	0.009	0.004	0.003	0.006	0.001	0.008	

The fusion methods evaluated are: i) Majority Voting (MV), ii) STAPLE [1], iii) COLLATE [18], iv) LWV [4], v) LOP-STAPLE [8], vi) empirical local MAP-STAPLE [13], and vii) our new optimal local MAP-STAPLE. The best Dice similarity results are marked in bold.

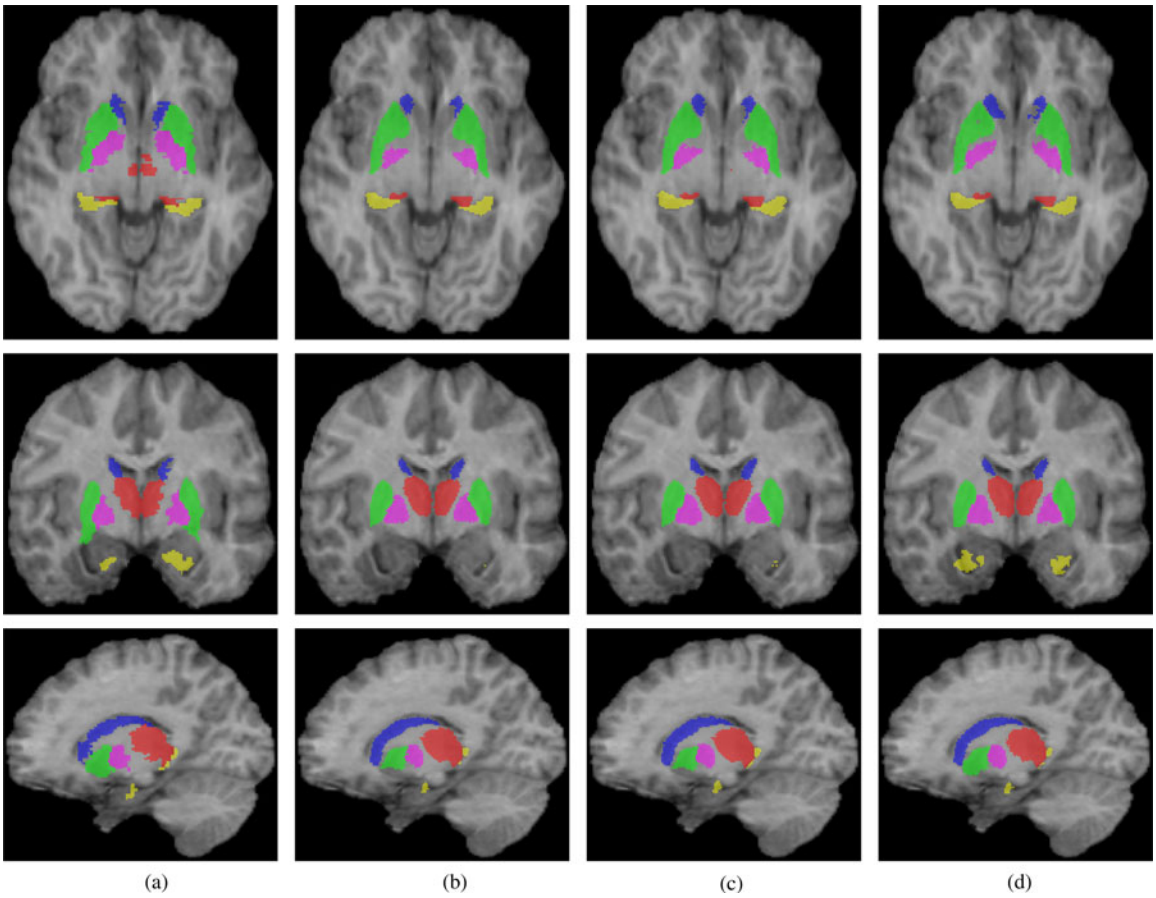

Fig. 1. Screen-shot of segmentation results for subcortical structures in one of images in the IBSR dataset. (a) Ground truth segmentations are shown in column; segmentation results from MV, STAPLE and optimal local MAP-STAPLE are shown in columns (b), (c), and (d), respectively. The segmentations for thalamus, caudate, putamen, pallidum and hippocampus are, respectively, shown in red, blue, green, magenta, and yellow, respectively. From qualitative comparisons with the ground truth segmentations in column (a), it can be noted that the proposed optimal local MAP-STAPLE has provided the best segmentation results among them. (a) Ground Truth. (b) MV. (c) STAPLE. (d) Opt. Local MAP-STAPLE.

Wilcoxon signed rank tests, namely, the p value, the sum of the ranks assigned to the differences with positive sign (V), and the confidence interval $[C_L\ C_H]$ associated with each comparison. Notice that, since these statistics are computed for ten structures, and for a dataset of 18 subjects, the maximum possible value of V is 16290. It is clear from the positive values of V that we got in all the six statistical tests, that, with 95% confidence, the Dice similarity coefficient values obtained from the proposed method are statistically better than the results from all the rest of the fusion methods. Thus, the results from the proposed method are found to be better than the other six methods, both quantitatively and statistically.

IV. Discussion and Conclusion

In this paper, we have presented a new approach for learning prior knowledge about the performance parameters of template images. We have also proposed a methodology for incorporating this prior knowledge into the STAPLE algorithm.

To the best of our knowledge, this is the first work that deals with learning prior knowledge about the performance parameters (i.e., sensitivity and specificity) *from the intensity information*. The prior knowledge about the performance parameters is inferred based on the local intensity similarity between each template image and the target image; it is then incorporated into the fusion method through the estimation of the optimal parameters of the MAP-based formulation of the STAPLE algorithm.

The proposed algorithm has been evaluated in the context of segmentation of structures in the brain MR images. We compared the proposed "optimal local MAP-STAPLE" algorithm with six state-of-the-art methods, namely, 1) MV, 2) STAPLE [1], 3) COLLATE [18], 4) LWV [4], 5) LOP-STAPLE [8], and 6) empirical local MAP-STAPLE [13].

Notice that among all the seven fusion methods, MV, STAPLE, COLLATE, and "local empirical MAP-STAPLE" algorithms do not take into account the intensity similarity information. On the other hand, voting-based fusion algorithms (MV and LWV), unlike the STAPLE-based algorithms, are not based on the explicit evaluation of rater performance parameters. In this perspective, among all the seven methods, the proposed algorithm and the LOP-STAPLE are the only methods that take into account both the intensity information, and the rater performance parameters.

When compared to the LOP-STAPLE algorithm, our proposed method uses the local intensity information in a very different manner. Notice that the LOP-STAPLE algorithm incorporates the local intensity information by modifying the way the reliability weights for each rater are computed in the EM-based STAPLE algorithm; on the contrary, our proposed algorithm learns prior knowledge about the performance parameters of each rater using the local intensity information, and then, incorporates it into the fusion process through the computation of optimal parameters of the MAP-based STAPLE algorithm. In addition, unlike the EM-based formulation, the MAP-based formulation used in our algorithm could result in convergence to a global optimum (or strong local optimum), and thereby, resulting in more accurate estimation of both the performance parameters and the output segmentations.

The aforementioned theoretical differences that we observe between different fusion methods are in coherence with the quantitative results obtained in the context of segmentation of structures in the brain MR images. For instance, LOP-STAPLE and the proposed fusion algorithm have provided the best segmentation results among all the methods; within those two methods, the proposed method has provided the best overall segmentation results. The improvements in the Dice similarity coefficient for the proposed method are found to be statistically significant when compared to the rest of the fusion methods.

As mentioned in the preceding section, the parameters for each fusion method (shown in Table I) are optimized independently. For the current application, our proposed algorithm is found to be robust and has less sensitivity to changes in theses parameters in a broad range. For instance, we observed that the behavior of the proposed algorithm to changes in scale (A) and shift (b) parameters is very similar to the sensitivity analysis results presented for the LOP-STAPLE algorithm in [8].

Similarly, in all our evaluations, we have empirically set the variance (σ^2) of the beta distribution to a fixed value of $1e{-}4$. Notice that setting variance to very high values is, in effect, equivalent to assuming a uniform prior regarding the performance parameters; on the contrary, very low values of variance force the algorithm to strictly converge to the prior knowledge that we have learnt based on the intensity information. In other words, the variance value indicates our confidence on the prior knowledge. We observed that, for the current application, the segmentation results from the proposed methods are quite robust to changes in the variance values. In the future work, we would like to perform a detailed sensitivity analysis for various parameters of the proposed algorithm and also explore the possibilities of learning these parameters from the training data.

Unlike in our preliminary work [15] where we assumed a strict linear relationship between the mode values of the beta distribution (of the performance parameters), and the intensity similarity, we proposed here to assume a more general exponential-based relationship. In the future work, we would like to investigate further regarding other possible relationships that one could assume between the intensity-similarity and the performance parameters. For instance, one could perhaps learn those relationships using various deep learning techniques and then, incorporate that information into the STAPLE algorithm. One could also explore other possible strategies like Bayesian learning of MAP parameters.

In the current study, we have considered the binary segmentation problem. It is indeed possible to extend the proposed method to multilabel segmentation problem. For binary segmentation, the system of equations for the performance parameters (5) has analytical closed-form solution (7). For a multilabel problem, although the system of equations do not have a closed-form solution, and although it can be computationally more expensive than the binary segmentation, it still has a unique solution (called fixed point). In the future work, we would like to extend the current framework to multilabel segmentation problem and also develop computationally more efficient models for multi-label fusion.

REFERENCES

[1] S. Warfield, K. Zou, and W. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. Med. Imag.*, vol. 23, no. 7, pp. 903–921, Jul. 2004.

[2] A. Akhondi-Asl and S. Warfield, "Simultaneous truth and performance level estimation through fusion of probabilistic segmentations," *IEEE Trans. Med. Imag.*, vol. 32, no. 10, pp. 1840–1852, Oct. 2013.

[3] A. Asman and B. Landman, "Formulating spatially varying performance in the statistical fusion framework," *IEEE Trans. Med. Imag.*, vol. 31, no. 6, pp. 1326–1336, Jun. 2012.

[4] X. Artaechevarria and A. Munoz-Barrutia, "Combination strategies in multi-atlas image segmentation: Application to brain MR data," *IEEE Trans. Med. Imag.*, vol. 28, no. 8, pp. 1266–1277, Aug. 2009.

[5] M. Sabuncu, B. Yeo, K. Van Leemput, B. Fischl, and P. Golland, "A generative model for image segmentation based on label fusion," *IEEE Trans. Med. Imag.*, vol. 29, no. 99, pp. 1714–1729, Oct. 2010.

[6] H. Wang, J. Suh, S. Das, J. Pluta, C. Craige, and P. Yushkevich, "Multi-atlas segmentation with joint label fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 611–623, Mar. 2013.

[7] S. Gorthi, M. B. Cuadra, P.-A. Tercier, A. Allal, and J.-P. Thiran, "Weighted shape-based averaging with neighborhood prior model for multiple atlas fusion-based medical image segmentation," *IEEE Signal Process. Lette.*, vol. 20, no. 11, pp. 1034–1037, Nov. 2013.

[8] A. Akhondi-Asl, L. Hoyte, M. Lockhart, and S. Warfield, "A logarithmic opinion pool based STAPLE algorithm for the fusion of segmentations with associated reliability weights," *IEEE Trans. Med. Imag.*, vol. 33, no. 10, pp. 1997–2009, Oct. 2014.

[9] S. Gorthi, "Multi atlas fusion methods for medical image segmentation," Ph.D. dissertation, Dept. Elect. Eng., Ecole polytechnique fdrale de Lausanne, Lausanne, Switzerland, 2013.

[10] T. Rohlfing and C. R. Maurer, Jr., "Shape-based averaging," *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 153–161, Jan. 2007.

[11] M. Cardoso, K. Leung, M. Modat, J. Barnes, and S. Ourselin, "Locally ranked staple for template based segmentation propagation," in *Proc. Workshop Multi-Atlas Labeling Statistical Fusion*, 2011, vol. 25, pp. 25–26.

[12] O. Commowick and S. K. Warfield, "Incorporating priors on expert performance parameters for segmentation validation and label fusion: A Maximum A Posteriori STAPLE," in *Proc. Med. Image Comput. Comput.-Assisted Intervention*, 2010, vol. 6363, pp. 25–32.

[13] O. Commowick, A. Akhondi-Asl, and S. K. Warfield, "Estimating a reference standard segmentation with spatially varying performance parameters: Local MAP STAPLE," *IEEE Trans. Med. Imag.*, vol. 31, no. 8, pp. 1593–1606, Aug. 2012.

[14] B. Landman, A. Asman, A. Scoggins, J. Bogovic, F. Xing, and J. Prince, "Robust statistical fusion of image labels," *IEEE Trans. Med. Imag.*, vol. 31, no. 2, pp. 512–522, Feb. 2012.

[15] S. Gorthi, A. Akhondi-Asl, J.-P. Thiran, and S. Warfield, " Optimal MAP parameters estimation in STAPLE—Learning from performance parameters versus image similarity information," in *Proc. Mach. Learn. Med. Imag.*, 2014, vol. 8679, pp. 174–181.

[16] S. M. AbouRizk, D. W. Halpin, and J. R. Wilson, "Visual interactive fitting of beta distributions," *J. Construction Eng. Manage.*, vol. 117, no. 4, pp. 589–605, 1991.

[17] A. J. Asman and B. A. Landman, "Characterizing spatially varying performance to improve multi-atlas multi-label segmentation," in *Information Processing in Medical Imaging*. New York, NY, USA: Springer, 2011, pp. 85–96.

[18] A. Asman and B. Landman, "Robust statistical label fusion through consensus level, labeler accuracy, and truth estimation (COLLATE)," *IEEE Trans. Med. Imag.*, vol. 30, no. 10, pp. 1779–1794, Oct. 2011.

[19] A. Klein, J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M.-C. Chiang, G. E. Christensen, D. L. Collins, J. Gee, P. Hellier, J. H. Song, M. Jenkinson, C. Lepage, D. Rueckert, P. Thompson, T. Vercauteren, R. P. Woods, J. J. Mann, R. V. Parsey, "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration," *Neuroimage*, vol. 46, no. 3, pp. 786–802, 2009.

[20] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Non-parametric diffeomorphic image registration with the demons algorithm," in *Proc. Med. Image Comput. Comput.-Assisted Intervention*, 2007, pp. 319–326.

[21] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache. (2007). Diffeomorphic demons using ITKs finite difference solver hierarchy. *Insight J.* [Online]. Available: http://www.insight-journal.org/browse/publication/154

Authors,' photographs and biographies not available at the time of publication.