Published in final edited form as:
IEEE Trans Med Imaging. 2013 October ; 32(10): 1840–1852. doi:10.1109/TMI.2013.2266258.

Simultaneous Truth and Performance Level Estimation Through Fusion of Probabilistic Segmentations

Alireza Akhondi-Asl[*] and **Simon K. Warfield**[*] **[Senior Member IEEE]**

Alireza Akhondi-Asl: Alireza.Akhondi-Asl@childrens.harvard.edu; Simon K. Warfield: Simon.Warfield@childrens.harvard.edu

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

Abstract

Recent research has demonstrated that improved image segmentation can be achieved by multiple template fusion utilizing both label and intensity information. However, intensity weighted fusion approaches use local intensity similarity as a surrogate measure of local template quality for predicting target segmentation and do not seek to characterize template performance. This limits both the usefulness and accuracy of these techniques. Our work here was motivated by the observation that the local intensity similarity is a poor surrogate measure for direct comparison of the template image with the true image target segmentation. Although the true image target segmentation is not available, a high quality estimate can be inferred, and this in turn allows a principled estimate to be made of the local quality of each template at contributing to the target segmentation. We developed a fusion algorithm that uses probabilistic segmentations of the target image to simultaneously infer a reference standard segmentation of the target image and the local quality of each probabilistic segmentation. The concept of comparing templates to a hidden reference standard segmentation enables accurate assessments of the contribution of each template to inferring the target image segmentation to be made, and in practice leads to excellent target image segmentation. We have used the new algorithm for the multiple-template-based segmentation and parcellation of magnetic resonance (MR) images of the brain. Intensity and label map images of each one of the aligned templates are used to train a local Gaussian mixture model based classifier. Then, each classifier is used to compute the probabilistic segmentations of the target image. Finally, the generated probabilistic segmentations are fused together using the new fusion algorithm to obtain the segmentation of the target image. We evaluated our method in comparison to other state-of-the-art segmentation methods. We demonstrated that our new fusion algorithm has higher segmentation performance than these methods.

I. Introduction

Multiple-template-based segmentation methods are widely used for the segmentation of medical images [1], [2], [3], [4], [5], [6], [7]. This type of algorithm has two distinct parts: non-rigid registration and fusion. Generally, in the first step, templates are registered to the target image and then the aligned templates and their corresponding label maps are used by a fusion algorithm to determine the segmentation of the target image.

The majority voting method is the simplest method for the fusion of the templates [8]. In this approach, at each voxel, the label with the largest number of votes is selected as the label of

the target image. One of the drawbacks of this approach is that all of the votes are equally weighted. However, because of the limitations of the registration algorithms and also the potential uncapturable inter-individual anatomical differences between some of the templates and the target image, the templates may not perform equally well. Thus, to have a more accurate result, the performance of the templates should be considered in the fusion process. The challenge here is that the performance of the templates is not known a-priori and needs to be estimated.

In the STAPLE algorithm [9], an Expectation-Maximization (EM) algorithm is used to simultaneously estimate the hidden ground truth and to estimate the performance of the aligned templates. At each iteration of the EM, performances of the templates are estimated based on comparison to an evolving estimate of the reference standard segmentation and the new performance parameters are used to update the reference standard segmentation. There are a number of extensions of the STAPLE algorithm which use a similar methodology for the fusion [10], [11], [12], [13], [14], [7], [15].

The local intensity similarity of the target image and the templates are also employed to estimate the local weight of the templates in the locally weighted fusion methods [1], [2], [16], [3], [4], [6], [17]. In this approach, the aligned label maps corresponding to the templates with higher intensity similarity to the target image have more weights in the fusion process. While in STAPLE based methods, performance of the templates are directly estimated by estimating the ground truth, there has been no direct association between the intensity similarity of the template and the target image, and the performance of the templates in the locally weighted fusion algorithms.

Recently, using intensity images of the templates in the fusion process have been considered in two STAPLE-based algorithms [18], [19]. The Multi-STEPS algorithm uses local intensity similarity of the templates and the target image to locally select the templates in the STAPLE algorithm [18]. Non-local-STAPLE algorithm uses local similarity of the templates and the target image in the STAPLE algorithm based on a non-local means approach to improve the segmentation accuracy [19].

To further improve the segmentation accuracy, two independent approaches can be employed: Improvement of the fusion strategy or improvement of the input segmentations fed to the fusion algorithm. To improve the fusion strategy, we need to estimate the template performance more accurately. This is the main feature of any fusion algorithm. It is obvious that even very powerful fusion strategies cannot accurately estimate the segmentation of the target image, if the information provided by the input segmentations is inadequate. Thus, improving the accuracy of the input segmentations generated from the templates will increase the accuracy of the fusion algorithm.

In many fusion algorithms, the aligned template segmentations have been used directly to infer the segmentation of the target image. Thus, one way to improve the accuracy of input segmentations is to use a more accurate registration algorithm in the alignment phase. However, to have a smooth and invertible transformation, a diffeomorphic image registration algorithm should be used which means the registration algorithm cannot in principle capture non-diffeomorphic inter-individual differences between the templates and the target image. Moreover, global optimization of the registration parameters is not feasible in a realistic registration problem and only a local optimum can be achieved. Furthermore, with a finite number of templates, it is not feasible to sample all anatomical variabilities and to feed the fusion algorithms with the information required to get the perfect segmentation of the target image. We propose instead to improve the target segmentation by using intensity and label map information. To this end, intensity and label map images of each one

of the aligned templates are used to train a classifier. We use a local Gaussian mixture model (GMM) as the classifier. Each one of the trained classifiers is used to segment the target image. The output of each one of the classifiers is a probabilistic segmentation of the target image.

The generated probabilistic segmentations could be the input segmentations of STAPLE as the desired fusion algorithm. However, the STAPLE algorithm cannot be used for the fusion of the probabilistic segmentations as it requires a segmentation label at each voxel. One way to solve this problem is to use the label with the highest probability at each voxel. However, we may lose some accuracy in this process. The challenging question is how these probabilistic segmentations can be directly fed into a fusion framework to simultaneously estimate the hidden ground truth and the performance parameters of the segmentation generators. In [20] an algorithm was proposed and validated that achieved for the first time, effective performance weighted fusion of probabilistic labels. The introduced algorithm, called soft-STAPLE, simultaneously estimates the performance of the segmentation generators and a reference standard segmentation from a collection of probabilistic segmentations of an image using an EM algorithm.

In this manuscript, we introduce a novel extension of the STAPLE algorithm, which uses an EM framework to simultaneously estimate the performance of the segmentations and the hidden ground truth from a collection of probabilistic segmentations which is computationally comparable with the original STAPLE algorithm. In general, the new fusion algorithm can be used for the fusion and evaluation of statistical classifiers, manual segmentations specified as confidence levels, or any set of probabilistic segmentations of a target image.

In this paper, we use our new fusion framework for the parcellation and segmentation of brain structures. We show that by using this approach, the performance of fusion is significantly improved compared to the classic STAPLE. In addition, we have compared our method to other state-of-the-art segmentation algorithms and have shown the superiority of our approach. The main contributions of our work are: development of a new fusion algorithm for the estimation of the reference standard segmentation and the performance parameters of segmentation generators from a set of probabilistic segmentations and estimation of probabilistic labeling using the intensity and label map images of the templates and the target image. The rest of the paper is organized as follows: Section II describes our novel fusion algorithm. Section II-F explains the proposed method for the calculation of probabilistic decisions based on the intensity and label information. Experimental data, procedures, and results are explained in Section III and conclusions are presented in Section IV.

II. Materials and Methods

In this part of the manuscript, we first describe our algorithm for the fusion of probabilistic decisions. Then, in Section II-F, we describe locally weighted fusion algorithms and our framework for the generation of probabilistic decisions using intensity and label information of the templates and the target image. Finally, we describe the imaging data utilized for the validation of our developed algorithm.

A. Probabilistic Fusion Algorithm

In this section, we developed our formulation for the fusion of probabilistic decisions. For this purpose, we first define the notations that will be used throughout the text. We assume that the goal is to estimate the hidden ground truth $\mathbf{T} = \{T_1, \ldots, T_i, \ldots, T_N\}$ of the target image where N is the number of voxels in the image and T_i indicates the label of the target

image at voxel i. Also, it is assumed that there are J independent raters where each one of them makes its own decision for each voxel. To estimate the hidden ground truth, we want to fuse the decisions made by the raters. We assume that each rater uses observable information to make the decision. We denote the observable information as $\mathbf{G} = \{\mathbf{G}_1,\ldots, \mathbf{G}_i, \ldots, \mathbf{G}_N\}$ where $\mathbf{G}_i = \{\mathbf{G}_{i1},\ldots, \mathbf{G}_{ij},\ldots, \mathbf{G}_{iJ}\}$ and $\mathbf{G}_{ij}$ is the observation used by the rater j to make the decision at voxel i. In addition, we denote the decisions as $\mathbf{D} = \{\mathbf{D}_1,\ldots, \mathbf{D}_i,\ldots, \mathbf{D}_N\}$ where $\mathbf{D}_i = \{D_{i1},\ldots,D_{ij},\ldots,D_{iJ}\}$ and D_{ij} is the decision of the rater j at voxel i. Next, we define Π as the probability map of decision by raters where $\Pi = \{\Pi_1, \ldots, \Pi_i,\ldots, \Pi_N\}$ and $\Pi_i = \{\Pi_{i1},\ldots,\Pi_{ij},\ldots, \Pi_{iJ}\}$. Here, $\Pi_{ij} = \{\pi_{ij0},\ldots, \pi_{ijs}, \ldots, \pi_{ijS}\}$ where $\pi_{ijs} = P(D_{ij} = s|\mathbf{G}_{ij})$ is the probability of the decision made by rater j at voxel i for label s, given the observation vector $\mathbf{G}_{ij}$. It is assumed that $\forall j, i : \Sigma_s \pi_{ijs} = 1$. In other words, rater j uses information in the observation $\mathbf{G}_{ij}$ to compute the probabilistic decision at voxel i.

In order to consider the variability of the rater performances, we consider $\theta = \{\theta_1,\ldots, \theta_j,\ldots, \theta_J\}$ as the descriptor of the agreement between the raters and the unknown ground truth. In addition, we define performance parameter $\theta_{js's} = P(D_{ij} = s'|T_i = s)$ as the elements of $S \times S$ matrix θ_j where $s, s' \in \{0,1,\ldots, S-1\}$ are the labels of ground truth and the decision by rater j at voxel i. Finally, the probability density function of the complete data can be defined as $P(\mathbf{G}, \mathbf{T}|\theta, \Pi)$. We are interested to find the unknown performance parameters of each rater and the hidden ground truth. The performance parameters of each rater are estimated by maximization of the complete data log likelihood function. Since the ground truth is not an observable variable, the EM algorithm may be utilized for computation of the conditional expectation of the complete data log likelihood function at the E-step and then its maximization based on the estimation of the performance parameters at the M-step.

$$Q(\theta|\theta^t)=E[\log f(\mathbf{G},\mathbf{T}|\theta,\mathbf{\Pi})|\mathbf{G},\theta^t,\mathbf{\Pi}]=\sum_i\sum_{T_i}\log f(\mathbf{G}_i,T_i|\theta,\mathbf{\Pi}_i)f(T_i|\mathbf{G}_i,\theta^t,\mathbf{\Pi}_i) \quad (1)$$

where θ^t denotes the previous estimation of θ at iteration t. We show at the end of this section how to model a spatial correlation between reference standard segmentation labels.

B. E-Step

In this step, the estimation of the hidden ground truth is derived given the estimated performance parameters (θ^t). Using Bayes rule, the conditional probability of the ground truth at the voxel i for label s can be written as:

$$f(T_i=s|\mathbf{G}_i,\theta^t,\mathbf{\Pi}_i)=\frac{\left[\prod_j f(\mathbf{G}_{ij}|T_i=s,\theta_j^t,\mathbf{\Pi}_{ij})\right] f(T_i=s)}{\sum_{s'}\left\{\left[\prod_j f(\mathbf{G}_{ij}|T_i=s',\theta_j^t,\mathbf{\Pi}_{ij})\right] f(T_i=s')\right\}} \quad (2)$$

We have considered the fact that raters are independent which means $f(\mathbf{G}_i|T_i=s,\theta^t,\mathbf{\Pi}_i)=\prod_j f(\mathbf{G}_{ij}|T_i=s',\theta_j^t,\mathbf{\Pi}_{ij})$. Next, we expand $f(\mathbf{G}_{ij}|T_i=s,\theta_j^t,\mathbf{\Pi}_{ij})$ over all the possible decisions using the following equation:

$$f(\mathbf{G}_{ij}|T_i=s,\theta_j^t,\mathbf{\Pi}_{ij})=\sum_{s''} f(\mathbf{G}_{ij}|D_{ij}=s'',T_i=s,\mathbf{\Pi}_{ij})f(D_{ij}=s''|T_i=s,\theta_j^t) \quad (3)$$

We assume that if the decision by a rater is known for voxel i, the computation of the probability of T_i does not depend to the observation $\mathbf{G}_{ij}$. In other words, $P(T_i|\mathbf{G}_{ij}, D_{ij}, \Pi_{ij}) = P(T_i|D_{ij}, \Pi_{ij})$ which shows conditional independency of T_i and $\mathbf{G}_{ij}$ when D_{ij} is known. Thus

$$f(T_i|\mathbf{G}_{ij}, D_{ij}, \mathbf{\Pi}_{ij}) = \frac{f(T_i, \mathbf{G}_{ij}, D_{ij}|\mathbf{\Pi}_{ij})}{f(\mathbf{G}_{ij}, D_{ij}|\mathbf{\Pi}_{ij})} = \frac{f(T_i, D_{ij}|\mathbf{\Pi}_{ij})}{f(D_{ij}|\mathbf{\Pi}_{ij})} \quad (4)$$

We can clearly see that:

$$\frac{f(\mathbf{G}_{ij}, T_i, D_{ij}|\mathbf{\Pi}_{ij})}{f(T_i, D_{ij}|\mathbf{\Pi}_{ij})} = \frac{f(\mathbf{G}_{ij}, D_{ij}|\mathbf{\Pi}_{ij})}{f(D_{ij}|\mathbf{\Pi}_{ij})} \quad (5)$$

This means that $f(\mathbf{G}_{ij}|D_{ij} = s'', T_i = s, \mathbf{\Pi}_{ij}) = f(\mathbf{G}_{ij} | D_{ij} = s'', \mathbf{\Pi}_{ij})$. When we do not have any prior knowledge about the decision of the raters, a uniform distribution for the decisions is the appropriate assumption. Thus, using this assumption and $f(D_{ij} = s''|\mathbf{G}_{ij}, \mathbf{\Pi}_{ij}) = \pi_{ijs''}$, and given $f(D_{ij}=s''|T_i=s, \theta_j^t)=\theta_{js''s}$ it can be seen that:

$$f(\mathbf{G}_{ij}|D_{ij}=s'', \mathbf{\Pi}_{ij}) = \frac{f(D_{ij}=s''|\mathbf{G}_{ij}, \mathbf{\Pi}_{ij}) f(\mathbf{G}_{ij})}{f(D_{ij}=s'')} = C_{ij}\pi_{ijs''} \quad (6)$$

where C_{ij} is a constant which depends on the prior on $f(\mathbf{G}_{ij})$ and $f(D_{ij})$. Thus, Eq. 3 can be rewritten as $f(\mathbf{G}_{ij}|T_i=s, \theta_j^t, \mathbf{\Pi}_{ij}) = C_{ij}\sum_{s''}\pi_{ijs''}\theta_{ijs''s}$. Using this relation, the weight variable W_{si}^t can be computed as:

$$W_{si}^t = f(T_i=s|\mathbf{G}_i, \theta^t, \mathbf{\Pi}_i) = \frac{\left[\prod_j (\sum_{s''}\pi_{ijs''}\theta_{js''s}^t)\right]\left[\prod_j C_{ij}\right] p_{si}}{\sum_{s'}\left[\prod_j \left(\sum_{s''}\pi_{ijs''}\theta_{js''s'}^t\right)\right]\left[\prod_j C_{ij}\right] p_{s'i}} = \frac{\left[\prod_j (\sum_{s''}\pi_{ijs''}\theta_{js''s}^t)\right] p_{si}}{\sum_{s'}\left[\prod_j \left(\sum_{s''}\pi_{ijs''}\theta_{js''s'}^t\right)\right] p_{s'i}} \quad (7)$$

where $p_{si} = f(T_i = s)$.

This estimate of the true segmentation is independent of the neighboring voxels. However, in many applications there are relationships between spatially related voxels where considering these relations improves the accuracy of the estimation of hidden ground truth. To consider these relations, a Markov random field (MRF) model is used. It is known that each voxel is related to other voxels in a local neighborhood. To model this relationship, the approach in [9] is used, where at each iteration the estimated ground truth is set as the initialization of the following equation:

$$\overline{W}_{si}^t \leftarrow \frac{1}{Z}\exp\left(\ln p_{si} + \ln\prod_j\left(\sum_{s'}\pi_{ijs'}\theta_{js's}^t\right) + \sum_m\sum_n J_{sn}\overline{W}_{nm}^t\right) \quad (8)$$

J_{sn} is the weight parameter which describes the spatial compatibility of the labels s and n, Z is the normalization constant and $m \in \{1, \ldots, N\}$. It should be mentioned that at each iteration, $\overline{W}_{si}^t$ is normalized using the constraint $\sum_s \overline{W}_{si}^t = 1$. The fixed point update equation is iterated to convergence which is guaranteed by the Brouwer fixed-point theorem [21]. After convergence, $\overline{W}_{si}^t$ is the updated version of estimated ground truth W_{si}^t using the mean field estimation. J_{sn} can be learned from the template data, as in [22], [23].

C. M-Step

With W_{si}^t, the conditional probability of label s being the true segmentation at voxel i, the expert performance can be estimated. To this end, using the fact that $f(\mathbf{G}_i, T_i|\theta, \mathbf{\Pi}_i) = [\prod_j f(\mathbf{G}_{ij}|T_i, \theta_j, \mathbf{\Pi}_{ij})]\, f(T_i)$, we can write:

$$\begin{aligned}Q(\theta|\theta^t)&=\sum_i\sum_s\log f(\mathbf{G}_i,T_i\\&=s|\theta,\mathbf{\Pi}_i)W_{si}^t\\&=\sum_i\sum_s\sum_j\log(f(\mathbf{G}_{ij}|T_i\\&=s,\theta_j,\mathbf{\Pi}_{ij}))W_{si}^t\\&+\sum_i\sum_s\log(p_{is})W_{si}^t=\sum_i\sum_s\sum_j\log\left(C_{ij}\sum_{s'}\pi_{ijs'}\theta_{js's}\right)W_{si}^t\\&+\sum_i\sum_s\log(p_{is})W_{si}^t\\&=\sum_i\sum_s\sum_j\log\left(\sum_{s'}\pi_{ijs'}\theta_{js's}\right)W_{si}^t\\&+\sum_i\sum_s\sum_j\log(C_{ij})W_{si}^t\\&+\sum_i\sum_s\log(p_{is})W_{si}^t\end{aligned}\tag{9}$$

The second and third terms on the right of the last equation do not depend on the parameters being optimized. Based on Jensen's inequality, $\log(\sum_{s'}\pi_{ijs'}\theta_{js's}) \geq \sum_{s'}\pi_{ijs'}\log(\theta_{js's})$. Thus, we just need to optimize:

$$Q'(\theta|\theta^t)=\sum_i\sum_s\sum_j\sum_{s'}\pi_{ijs'}\log(\theta_{js's})W_{si}^t \tag{10}$$

Finally, using the constraint $\sum_{s'}\theta_{js's}=1$, it is possible to update performance parameter $\theta_{js's}$ using the following equation:

$$\theta_{js's}^{t+1}=\frac{\sum_i\pi_{ijs'}W_{si}^t}{\sum_{n'}\sum_i\pi_{ijn'}W_{si}^t}=\frac{\sum_i\pi_{ijs'}W_{si}^t}{\sum_i W_{si}^t} \tag{11}$$

D. MAP Solution

It is possible to find the maximum a posteriori solution of the problem when a prior about the performance parameters are known. In [12] authors have explained a practical solution for this problem for the classic STAPLE fusion. Let $P(\theta)=\prod_{j,s,s'}P(\theta_{js's})$ be the probability map of the prior and γ be the weight of the prior term. Thus, we can write the map formulation as:

$$Q_{MAP}(\theta|\theta^t)=Q'(\theta|\theta^t)+\gamma\log(P(\theta))=Q'(\theta|\theta^t)+\gamma\sum_j\sum_s\sum_{s'}\log(P(\theta_{js's})) \tag{12}$$

Using the beta distribution with parameters $\alpha_{js's}$ and $\beta_{js's}$ for each rater j and labels s and s', the prior probabilities are set to $P(\theta_{js's})=\frac{1}{Z}(\theta_{js's})^{\alpha_{js's}-1}(1-\theta_{js's})^{\beta_{js's}-1}$ where Z is the normalization constant. Thus, using the same procedure above, it can be shown that:

$$Q_{MAP}(\theta_j|\theta^t)=\gamma\sum_{s'}\sum_s\left((\alpha_{js's}-1)\log(\theta_{js's})+\beta_{js's}-1)\log(1-\theta_{js's})\right)+\sum_i\sum_s\sum_{s'}\pi_{ijs'}W_{si}^t\log(\theta_{js's}) \tag{13}$$

Optimization of this function with the constraint $\sum_{s'}\theta_{js's} = 1$, leads to the following equation for each one of the performance parameters:

$$\theta_{js's}^{t+1}=\frac{(\sum_i \pi_{ijs'} W_{si}^t)+\gamma(\alpha_{js's}+\beta_{js's}-2)+\gamma\frac{\beta_{js's}-1}{\theta_{js's}^{t+1}-1}}{\sum_{n'}\left[(\sum_i \pi_{ijn'} W_{si}^t)+\gamma(\alpha_{jn's}+\beta_{jn's}-2)+\gamma\frac{\beta_{jn's}-1}{\theta_{jn's}^{t+1}-1}\right]} \quad (14)$$

Generally, there is no closed form solution for this equation. However, $\theta_j = f(\theta_j)$ with $f:]0, 1[^{s^2} \leftarrow]0, 1[^{s^2}$ always has a unique fixed point solution. The solution can be achieved through iterative application of the equation by commuting the sequence $\{x_n\}_{n\geq 1}$ where $x_{n+1} = f(x_n)$ until convergence. A closed form solution is available for all priors with $\beta_{js's} = 1$ and this can be used to provide an initial estimate for the iterative estimator for $\beta_{js's} \neq 1$.

E. Relation to STAPLE and Local MAP STAPLE

We have developed a formulation for the estimation of the performance parameters and ground truth simultaneously. In this formulation, we have used probabilistic decision of the raters for the estimation. Classic STAPLE is designed for the estimation of the performance parameters and ground truth when the decisions are crisp. It is easy to see the relationship between our developed formulation and the classic STAPLE. For this purpose, it should be noted that in the STAPLE formulation, decision of rater j at voxel i is the label of template at voxel i. Let $\mathbf{L} = \{\mathbf{L}_1,\ldots, \mathbf{L}_i,\ldots, \mathbf{L}_N\}$ be the matrix of size $N \times J$ with the vectors $\mathbf{L}_i = [L_{i1},\ldots, L_{ij}, \ldots, L_{iJ}]'$ where L_{ij} is the label of the template at voxel i. In the classic STAPLE, $\mathbf{G}_{ij} = \{L_{ij}\}$ and $\pi_{ijs} = \delta(L_{ij} - s)$ where $\delta(x)$ is the Kronecker delta function. This means that $\sum_{s'}\pi_{ijs'}\theta_{js's}^t = \theta_{jL_{ij}s}^t$ and $\sum_i \pi_{ijs'} W_{si}^t = \sum_{i:L_{ij}=s'} W_{si}^t$ where substitutions of these results to Eqs. 7, 11, and 14 will lead to the classic STAPLE formulation. Thus, our formulation is a generalized formulation of the STAPLE algorithm as in the special case of the crisp decisions, our formulation will be equivalent of the STAPLE formulation.

Furthermore, the above formulation computes the performance parameters globally. However, it is known that the performance of each one the templates may vary locally based on the accuracy of the registration of the template segmentation and also topology of the templates [7]. Local MAP STAPLE has been introduced to take into account this point in the fusion process [7]. The method estimates a reference standard and spatially varying performance parameters for each one of the templates in local regions. To do this, for each voxel we define a region of a specified known half window size around it. Then, we run STAPLE using the voxels in the region only to find the local performance parameters and local reference standard segmentation. $\mathbf{T}$ and θ for a region are defined to be parameters of the central voxel. Finally, we slide the window one voxel and use the same procedure to cover all of the voxels in the image. In this way, performance of each one the templates is estimated locally. A potential problem with this approach is that for the templates, labels of one or more structures could be missing in some windows. The absence of some structures may cause STAPLE to converge to an incorrect local maximum as be unable to estimate the performance parameters. This can be avoided by using the prior knowledge of the expert performance parameters, using MAP parameters as described in the previous section [12], [7].

In summary, we have introduced our novel probabilistic fusion algorithm which simultaneously estimates the hidden ground truth and performance of the raters. We have presented a general probabilistic fusion framework which has wide range of applications including fusion and evaluation of statistical classifiers, manual segmentations specified as confidence levels, or any set of probabilistic decisions. In the next section, we first describe

locally weighted fusion methods which use local intensity similarity of the templates and the target image in the fusion process and then, we explain our approach for generation of the probabilistic decisions using the intensity and label information of the templates and the target image.

F. Using Intensity and Label Information in the Fusion Process

In the previous section, we introduced our fusion algorithm which simultaneously estimates the performance parameters of the segmentations and the hidden ground truth. We showed that the described algorithm is a generalized formulation of the STAPLE algorithm which works with probabilistic decisions. In this section, we will discuss locally weighted fusion algorithms for the segmentation and introduce our framework for the generation of probabilistic decisions using intensity and label information of the templates and the target image. Similar to the previous section, we require that there are J templates and their corresponding segmentations. For simplicity, we also assume that the templates are aligned to the target image using a non-rigid registration algorithm and are independent of each other. Let U_{ij} be the intensity of the aligned template j at voxel i. Let $\mathbf{U} = \{\mathbf{U}_1, \ldots, \mathbf{U}_i, \ldots, \mathbf{U}_N\}$ be the matrix of size $N \times J$ with the vectors $\mathbf{U}_i = [U_{i1}, \ldots, U_{ij}, \ldots, U_{iJ}]'$. In addition, we assume that $\mathbf{I} = \{I_1, \ldots, I_i, \ldots, I_N\}$ is the target image where I_i shows the intensity of the target image at voxel i.

1) Intensity Similarity Based Locally Weighted Fusion Algorithms : Fusion of segmentations can be done using different strategies. Majority voting approach is the simplest approach that can be used for this purpose. In this approach, decision of rater j at voxel i is equal to L_{ij}, the label of the template at the voxel i. The main problem of this approach is that it considers all of the raters equally. To improve the performance of the fusion, in [24] authors have suggested using signed distance map of the segmentations rather than the binary decisions. The idea is that the segmentations are more accurate for the points that are deep inside the structures and for the points close to the boundary of the structures, the template segmentations are less accountable. Same as the majority voting, raters are considered equally weighted and the intensity information is not considered in the fusion process.

Recently, intensity information has been used in the fusion process with the goal of improvement of segmentation estimation. Almost all of these intensity similarity based locally weighted fusion methods can be described using the following formulation:

$$f(T_i = s|\mathbf{U}, \mathbf{I}, \mathbf{L}) = \frac{\sum_j f(I_{\mathcal{N}\{i\}}|U_{\mathcal{N}\{i\}j}) f(D_{ij} = s|L_{.j})}{Z} = \frac{\sum_j w_{ij} f(D_{ij} = s|L_{.j})}{Z} \quad (15)$$

where $[L_{.j} = \{L_{ij} | \forall i \in \{1, \ldots, N\}\}$ and $\mathcal{N}i\}$ is a local region around voxel i. In this formulation w_{ij} is the weight of rater j at voxel i which is defined to depend on intensity similarity of the target image and the template. The idea is to give a larger weight to the templates with higher similarity to the target image. The limitation of this approach is that there is no clear link between the weights and the segmentation quality as they maybe associated with poor registration or unimportant intensity variations. However, STAPLE based methods which utilize intensity information of the templates in the fusion process minimize this problem, since they directly estimate the performance of the templates and intensity similarity is not the only source of performance estimation.

A variety of methods have been used in the literature to estimate this weight. In one of the first attempts, Artaechevarria et al. compared a number of metrics such as mean square error, mutual information, and normalized cross correlation to find the best metric to estimate the weight [1]. They showed that mean square error is the best metric for this

purpose. Also, they investigated effect of the neighborhood radius on the fusion results. This strategy was also used for heart segmentation [16]. Sabuncu et al. used a Gaussian kernel as the similarity metric [2]. Yushkevich et al. sorted all of the templates based on their similarities and then used ranks to weight different templates [4]. Regression models have also been used to find the weights [17], [6]. Khan et al. developed a supervised method where the training datasets were used to estimate the weight model parameters [6]. Wang et al. utilized intensity similarity between the target image and each pair of template images to estimate the weight of the templates [25] . In this way, voxelwise intensity based weights are not estimated independently and the correlation between the segmentation errors is considered in the process of voxelwise weight estimation.

In order to allow us to experimentally compare to such methods, we have implemented three configurations of the locally weighted fusion formulations [1], [2], [25]. In [1] the following fusion strategy is used:

$$f(T_i=s|\mathbf{U},\mathbf{I},\mathbf{L})=\frac{\sum_j\left(\sum_{n\in\mathcal{N}\{i\}}(I_n-U_{nj})^2\right)^{-\gamma}\delta(L_{ij}-s)}{Z} \quad (16)$$

In this formulation, for each template j the decision at voxel i, L_{ij}, is weighted by the inverse of the dissimilarity of the target image and the template j in the neighborhood $\mathcal{N}i\}$. The dissimilarity is estimated using mean square differences of the intensities. Also, γ is a parameter of the method that can be set to control the sensitivity of the weights to the intensity differences. In another work, the weights are estimated using a Gaussian kernel with a fixed σ for all of the templates [2]:

$$f(T_i=s|\mathbf{U},\mathbf{I},\mathbf{L})=\frac{\sum_j\left(\frac{1}{\sqrt{(2\pi\sigma^2)}}\exp\left(-\frac{\sum_{n\in\mathcal{N}\{i\}}(I_n-U_{nj})^2}{2|\mathcal{N}\{i\}|\sigma^2}\right)\right)\delta(L_{ij}-s)}{Z} \quad (17)$$

Finally, methods of [25] uses the pairwise dependency matrix $\mathbf{M}_i$ for each voxel i to find the weights as the solution of the following minimization problem:

$$\min_{\mathbf{w}_i}\mathbf{w}_i\mathbf{M}_i\mathbf{w}_i+\alpha\|\mathbf{w}_i\| \; subject\ to: \sum_j w_{ij}=1 \quad (18)$$

where $\mathbf{M}_i$ is the matrix of size $J\times J$ with the elements $\mathbf{M}_{ijj'}=(\sum_{n\in\mathcal{N}\{i\}}|I_n-U_{nj}||I_n-U_{nj'}|)^{-\beta}$, α is the optimization parameter, β is the parameter which controls the sensitivity of the weights to the intensity differences, and $\mathbf{w}_i=\{w_{i1}\ldots w_{iJ}\}$. The authors also presented a local search technique to improve the registration accuracy of the templates which can be used in any other fusion strategy.

2) Training Local GMM Based Classifier Using Intensity and Label Information : In the STAPLE formulation, the aligned label map images are considered as the input segmentations. However, since there is no single brain which can represent 3D anatomical variations of all brains and because of the changes in the topology and folding patterns of the brains and more importantly, because of the possible structural abnormalities, brain anatomy exhibits non-diffeomorphic variation [26], [27], [28], [29], [30]. Currently, no diffeomorphic registration algorithm captures all of the inter-individual anatomical differences between the templates and the target image. In addition, the number of the templates is limited which makes it impractical to sample all anatomical variability and thus to perfectly segment the target image using direct fusion of the aligned label maps [31], [32], [33], [34].

In this section, we describe our approach to use intensity and label information of the templates and the target image to overcome this limitation. To this end, each one of the aligned templates and the corresponding label maps are used to train a GMM based classifier which is used to segment the target image. In this way, using the intensity and label map information of each one of the templates, a modified segmentation of the target image can be achieved which is less sensitive to the errors due to uncaptureable inter-individual anatomical dissimilarities between the target image and the templates and also the errors due to intrinsic registration inaccuracies.

Let us define local region $\mathcal{N}i\}$ around voxel i. Then, we can build decision probabilities using the following equation:

$$\pi_{ijs}=f(D_{ij}=s|\mathbf{G}_{ij})=\frac{\omega_{ijs}K(I_i,\mu_{ijs},\sigma^2_{ijs})}{\sum_{s'}\omega_{ijs'}K(I_i,\mu_{ijs'},\sigma^2_{ijs'})} \quad (19)$$

where $\mathbf{G}_{ij}=\{L_{mj}, U_{mj}|m\in\mathcal{N}i\}\}$. In this equation, the mean and standard deviation of label s at voxel i are estimated based on the following equations:

$$\mu_{ijs}=\frac{\sum_{m\in\mathcal{N}\{i\}}\delta(L_{mj}-s)U_{mj}}{\sum_{m\in\mathcal{N}\{i\}}\delta(L_{mj}-s)}\ \sigma^2_{ijs}=\frac{\sum_{m\in\mathcal{N}\{i\}}\delta(L_{mj}-s)U^2_{mj}}{\sum_{m\in\mathcal{N}\{i\}}\delta(L_{mj}-s)}-\mu^2_{ijs}\ \omega_{ijs}=\frac{\sum_{m\in\mathcal{N}\{i\}:L_{mj}=s}\eta_{mi}}{\sum_{m\in\mathcal{N}\{i\}}\eta_{mi}} \quad (20)$$

Here, $K(I_i,\mu_{ijs},\sigma^2_{ijs})=\frac{1}{\sqrt{(2\pi\sigma^2_{ijs})}}\exp\left(-\frac{(I_i-\mu_{ijs})^2}{2\sigma^2_{ijs}}\right)$ and $\eta_{mi}=\mathcal{H}(\upsilon-d^2_{mi})$ where d_{mi} is the euclidean distance between voxels i and m, υ is a fixed tuning parameter, and $\mathcal{H}$ is the Heaviside function. It should be mentioned that for the cases that $\omega_{ijs}=0$, $\pi_{ijs}=0$.

For each template, π_{ijs} is a probabilistic segmentation of the target image which is used as the input to our new fusion algorithm. We refer to the fusion method which utilizes Eqs. 7,11, and 19 for the estimation of the ground truth and the performance parameters as PSTAPLE and refer to the fusion method which utilizes Eqs. 7,14, and 19 as Local MAP PSTAPLE.

It is also possible to fuse the probabilistic segmentations of the target image using the following probabilistic voting algorithm:

$$f(T_i=s|\mathbf{U},\mathbf{I},\mathbf{L})=\frac{\sum_j\pi_{ijs}}{\sum_j\sum_{s'}\pi_{ijs'}} \quad (21)$$

where π_{ijs} is computed using the Eq. 19.

In the next section, we evaluate our method using both synthetic data and brain images. To show that both of our novel fusion algorithm and the generated probabilistic segmentations are the source of improvement in the segmentation quality, we also compare our algorithm with the probabilistic voting algorithm described in Eq. 21.

G. Imaging Data

In the next section, we validate our algorithm using the synthetic data, the database of the brain MRI of the Internet Brain Segmentation Repository (IBSR), and the Neuromorphometric database (NMM). For the synthetic data, we generated eight synthetic images and their corresponding segmentations of the target image as shown in Fig. 1. Three of the templates are generated using the pattern in Fig. 1.a where their corresponding segmentation is shown in Fig. 1.d. Also, three templates are generated in the same manner

using the pattern shown in Fig. 1.b and their corresponding segmentation in Fig. 1.e. Finally, two remaining templates are based on the pattern in Fig. 1.c with their corresponding manual segmentation in Fig. 1.f. The target image and the ground truth are shown in Fig. 1.g and Fig. 1.h, respectively.

IBSR database includes 18 1.5T T1-weighted volumetric images with slightly different voxel sizes and their corresponding manual segmentation. 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/. Manual segmentation of 34 principle gray and white matter structures of the brain (3rd Ventricle, 4th Ventricle, Brain Stem, CSF and Left and Right: Accumbens area, Amygdala, Caudate, Cerebellum Cortex, Cerebral Cortex, Cerebellum White Matter, Cerebral White Matter, Hippocampus, Inf Lat Vent, Lateral Ventricle, Pallidum, Putamen, Thalamus Proper, VentralDC, and vessel) are available. In addition, the same database has the parcellation of 96 cortical gray matter parcels which makes the total number of 128 structures.

We collected and segmented whole brain MRI data in order to form an evaluation database for the comparison of segmentation algorithms. Fifteen healthy volunteers underwent MRI. We acquired T1-weighted; T2-weighted FSE (Fast Spin Echo); FLAIR-FSE; and diffusion weighted images of 15 volunteers on a 3T clinical MR scanner from GE Medical Systems (Waukesha, WI, USA) using an 8-channel receiver head coil. The T1-weighted images were acquired sagittally with a matrix size of 256 × 256 and a field of view of 24 cm. Slice thickness was around 1.3 mm and the T1-weighted acquisition parameters were TR 10/TE 6/TI 725 ms with a flip angle of 8. Each T1-weighted acquisition was carefully segmented into 53 structures by Neuromorphometrics, Inc. (http://www.neuromorphometrics.com/) using a well-established and carefully validated manual segmentation protocol [35], [36]. For three of the subjects, a second repeat manual segmentation was carried out in order to allow evaluation of test-retest reproducibility of the manual segmentation. The manually segmented 53 structures were: 3rd, 4th, and 5th Ventricles, Brain stem, CSF, Cerebellar Vermal Lobules I-V, VI-VII, and VIII-X, Optic Chiasm, and Left and Right: Accumbens area, Amygdala, Caudate, Cerebellum Hemisphere, Cerebellum White Matter, Hippocampus, Inf Lat Vent, Lateral Ventricle, Pallidum, Putamen, Thalamus Proper, VentralDC, vessel, and frontal, occipital, parietal, temporal lobe grey and white matter. We refer to this dataset as the NMM database, and use it for the evaluation of our segmentation algorithm, and the comparison of our algorithm to other state-of-the-art segmentation algorithms.

III. Results

In the previous section, we described our framework for the fusion of probabilistic segmentations where we have used a local GMM for the calculation of the probabilistic decisions. The estimated probabilities are based on the intensity and label information of the templates and the intensity information of the target image. In this section, we test and evaluate our segmentation method using synthetic data and two brain databases: the databases of IBSR and NMM.

A. Synthetic Data

In the first step, we utilize the synthetic data to compare our developed algorithm and the STAPLE algorithm. The segmentation result using PSTAPLE is similar to the ground truth (Fig. 1.h) and the segmentation results using the STAPLE is shown in Fig. 1.i. It can be seen that STAPLE does not achieve the correct segmentation as it is impossible to distinguish between the segmentations of the templates generated using the patterns in Fig. 1.d and Fig. 1.e. However, using local GMM based classifiers, which is employed to build the

probabilistic segmentation, our new fusion algorithm accurately estimates the target segmentation. In this simple example, we have shown the effectiveness of the modified segmentations in the fusion process. Next, we compare our method to the state-of-the-art fusion algorithms.

B. Methods used for the comparison and selecting their parameters

In the next section, we use two databases of the brain MRI and compare our developed method to the state-of-the-art methods in the literature. These methods are methods of [1], [2], [25] as the intensity similarity based locally weighted fusion methods, methods of [14], [9], [10], [11], [7] as the STAPLE based methods, and methods of [18], [19] as the STAPLE based methods which utilize intensity images of the templates in the fusion process. Since our method has two distinct components that can improve the segmentation accuracy, we also report the segmentation results using majority voting and voting using the probabilistic output of the GMM classifier. For the methods of [14], [10], [11] and [25] we have used the software packages given in https://masi.vuse.vanderbilt.edu/ and http://www.nitrc.org/projects/picslmalf/. Also, we have implemented methods of [1], [2], [19], [18]. In addition, for the STAPLE and Local MAP STAPLE algorithms, we have used the implementation given in http://crl.med.harvard.edu/software/STAPLE/.

We have used our methods, NLS, Multi-STEPS, and STAPLE with assigned consensus region which was first suggested in [37]. In this way, large consensus regions do not have any effect on the estimation of the performance parameters. In summary, we analyze the segmentation results of the following methods:

- $\mathcal{M}_0$: Majority Voting
- $\mathcal{M}_1$: Artaechevarria et al. method [1]
- $\mathcal{M}_2$: Sabuncu et al. algorithm [2]
- $\mathcal{M}_3$: SIMPLE [14]
- $\mathcal{M}_4$: STAPLER [11]
- $\mathcal{M}_5$: COLLATE [10]
- $\mathcal{M}_6$: STAPLE [9]
- $\mathcal{M}_7$: Multi-STEPS [18]
- $\mathcal{M}_8$: NLS [19]
- $\mathcal{M}_9$: PICSL-MALF [25]
- $\mathcal{M}_{10}$: Probabilistic Voting
- $\mathcal{M}_{11}$: PSTAPLE
- $\mathcal{M}_{12}$: Local MAP STAPLE [7]
- $\mathcal{M}_{13}$: Local MAP PSTAPLE

As discussed previously, STAPLE based methods do not use aligned intensity images of the templates in the fusion process. However, locally weighted fusion methods [1], [2], [25], NLS and Multi-STEPS [19], [18], and our proposed algorithm use these images in the fusion process. Thus, the intensities of the target image and the aligned templates are required to be matched using a normalization approach, otherwise the intensity images of the templates would not be suitable for the training of the classifiers. To solve this problem, one can use simple histogram matching, the methods described in [3], [17] or any other intensity normalization approach. We have used the combination of histogram matching and the

method described in [17]. In other words, for all of the consequent experiments a histogram matching method is used to globally match the intensity of the target image and the templates. Moreover, the same approach in [17] is used to locally normalize the intensities by using local mean and variance. In addition, we have used the leave-one-out method for the evaluation of the methods. We have examined different non-rigid registration algorithms and based on these experiments, we have used SyN for the alignment of the templates to the target image [38]. The output transformations were applied to the manual segmentation of the templates.

For the methods that do not use intensity images of the templates in the fusion process [14], [9], [10], [11], [7], we used their default parameters; however, using three cases from NMM dataset, we examined different parameters of [1], [2], [25], [18], [19]. The optimum HWS and σ for the method of [2] was 2 and 0.5 respectively and the optimum HWS and γ for the method of [1] was 2 and 1 respectively. The optimized parameters of PICSL-MALF were $HWS = 2$, $\beta = 2$, and $\alpha = 0.1$ [25]. In addition, the optimum HWS, local search radius, σ_i, and σ_d of the NLS was 2, 5, 0.1, and 1, respectively [19]. Finally, the optimized kernel parameter and number of templates of the method of Multi-STEPS was 2 and 6, respectively. For our method we used HWS of 5 and $\upsilon = 6$. It should be mentioned that the methods of [14], [9], [10], [11] do not perform well in the whole brain segmentation problem. The reason is that these methods use global performance parameters which is not appropriate for the whole brain segmentation [7]. The methods of [19], [18] have the same limitation; however, they use intensity images of the templates in the fusion process which moderately overcomes the problem. Thus, for the whole brain segmentation problem, we compared our method (Local MAP P STAPLE) with the methods of [1], [2], [25], [18], [19] and Local MAP STAPLE [7].

C. IBSR and NMM databases

In the first experiment, we have used each one of the 128 structures in the IBSR database and compared our method (PSTAPLE) to the other methods. Table I shows the Dice coefficient of $16 \times 2 + 1 = 33$ structures and the average Dice coefficient of 128 structures for the 12 methods. A paired-samples t-test was conducted to compare Dice coefficient of PSTAPLE and Multi-STEPS as the next most accurate method. There was a significant difference in the Dice coefficient of PSTAPLE ($\mu = 0.717$, $\sigma = 0.12$) and Multi-STEPS ($\mu = 0.710$, $\sigma = 0.12$); $t(127) = 2.4168$, $p < 0.05$, and 95% confidence interval of [0.001 0.012]. In addition, there is a statistically significant difference between PSTAPLE ($\mu = 0.717$, $\sigma = 0.12$) and Probabilistic voting ($\mu = 0.683$, $\sigma = 0.16$); $t(127) = 7.6437$, $p < 0.05$, and 95% confidence interval of [0.025 0.043]. Moreover, comparison of PSTAPLE ($\mu = 0.717$, $\sigma = 0.12$) and STAPLE ($\mu = 0.708$, $\sigma = 0.13$) shows a statistically significant improvement; $t(127) = 2.1234$, $p < 0.05$ and confidence interval of [0.001 0.018]. The results indicates that our new fusion algorithm has higher performance compared to the other fusion algorithms. Also, it can clearly be seen that both GMM classifier and fusion algorithm have statistically significant contribution in the improvement. To show the superiority of our method in the multi-category segmentation, we carried out two different experiments. In the first experiment, we used the generalized Dice coefficient to compare accuracy of Local MAP PSTAPLE and Local MAP STAPLE for the whole brain segmentation. Generalized Dice coefficient is computed as:

$$D = \frac{2\sum_l |A_l \cap B_l|}{\sum_l (|A_l| + |B_l|)} \quad (22)$$

where A and B indicate the two multi-label segmentations to compare and l denotes the label of the structures. Fig. 2 shows the comparison of two methods for the segmentation of 34

structures in the IBSR datasets. Average generalized Dice coefficient of our method is 0.933 while Local MAP STAPLE has the average of 0.912. Also, a paired-samples t-test was conducted to compare Dice coefficient of Local MAP PSTAPLE and Local MAP STAPLE. The Dice coefficient difference between local MAP PSTAPLE ($\mu = 0.933$, $\sigma = 0.0095$) and local MAP STAPLE ($\mu = 0.912$, $\sigma = 0.0148$) was statically significant; $t(17) = 12.9998$, $p < 0.05$ and confidence interval of [0.018 0.024]. It can be seen that our method is significantly superior to the Local MAP STAPLE.

In addition, we have compared our method with the methods of [1], [2], [25], [19], [18]. We used generalized Dice coefficient for the comparison and the results are shown in Fig. 3. Comparison of our method ($\mu = 0.934$, $\sigma = 0.0095$) and method of [25] ($\mu = 0.924$, $\sigma = 0.0115$) as the next most accurate method indicates a statistically significant difference between the methods; $t(17) = 9.5269$, $p < 0.05$ and confidence interval of [0.008 0.012]. It is also can be seen that while NLS and Multi-STEPS outperform locally weighted fusion methods in the binary segmentation problem, they have poor performance in the multi-category segmentation. Average generalized Dice coefficient of NLS and Multi-STEPS for IBSR dataset is 0.915 and 0.912, respectively. It is worth mentioning that local MAP implementation of these methods may overcome this problem and may significantly increase their accuracy.

In addition, Fig. 4 shows the comparison of the segmentation results generated by Local MAP PSTAPLE (f), method of [1] (a), method of [2] (b), Multi-STEPS (c), NLS (d), PICSL-MALF (e) and expert manual segmentation (g) in two representative IBSR dataset (h).

Next, we used our developed method (Local MAP PSTAPLE) and methods of [1], [2], [25], [19], [18] for the segmentation of NMM datasets. The same parameters described in the previous section are used in this experiment. Fig. 5 shows the comparison of the segmentation results generated by Local MAP PSTAPLE (f), method of [1] (a), method of [2] (b), Multi-STEPS (c), NLS (d), PICSL-MALF (e), and expert manual segmentation (g) in coronal images of two representative NMM datasets (h). In addition, quantitative evaluation with generalized Dice coefficient is also considered to compare the methods. The results of the comparisons are shown in the Fig. 6. The average generalized Dice coefficient of our method, method of [1], method of [2], Multi-STEPS, NLS, and PICSL-MALF are 0.851, 0.846, 0.846, 0.839, 0.833, and 0.844 respectively. The differences between our developed method ($\mu = 0.851$, $\sigma = 0.0169$) and the method of [1] ($\mu = 0.846$, $\sigma = 0.0150$) and method of [2] ($\mu = 0.846$, $\sigma = 0.0150$) as the next most accurate methods were statistically significant with $t(14) = 4.9536$, $p < 0.05$, and confidence interval of [0.003 0.007] and $t(14) = 5.3940$, $p < 0.05$ and confidence interval of [0.003 0.007], respectively. However, there was no difference between the methods of [1] and [2]; $t(14) = 1.4595$, $p > 0.05$ and confidence interval of [−0.0001 0.001].

IV. Conclusions

We have developed a new Expectation-Maximization algorithm that simultaneously estimates the hidden ground truth and performance parameters from a set of probabilistic segmentations. The algorithm can be used for the fusion and evaluation of statistical classifiers, manual segmentations specified as confidence levels, or any set of probabilistic segmentations of a target image.

In this paper, we have used the algorithm for the parcellation and segmentation of brain structures. Intensity and label map images of each one of the aligned templates have been used to train a local GMM based classifier. Each one of the trained classifiers has been used

to segment the target image which generates a probabilistic segmentation of the target image. We then locally estimate template performance by comparison to a hidden reference standard. In this way, we have identified local regions where the template matches or does not match due to uncaptureable inter-individual anatomical dissimilarities between the target image and the template and also the errors due to intrinsic registration inaccuracies which are then fused.

As a result of this work, we have introduced two distinct fusion algorithms: PSTAPLE, which uses global performance parameters and is useful when the performance of the raters does not vary spatially; and Local MAP PSTAPLE, which addresses the problem of spatial variability of the rater performance.

These new fusion algorithms were tested on two brain databases. These results, in turn, were compared to those of other state-of-the-art segmentation methods. We demonstrated that PSTAPLE and Local MAP PSTAPLE have higher accuracy compared to other state-of-the-art methods described in the literature.

The algorithm we have described enables the fusion of probabilistic input segmentations. Many intensity and label fusion algorithms generate such probabilistic segmentations. Our fusion approach allows the characterization of the performance of such algorithms even in the absence of a reference standard segmentation.

Although we have used a local GMM based classifier to generate the probabilistic segmentations, it may be possible to improve the accuracy of the segmentation by utilizing more sophisticated classifiers. This is one of the important aspects of the algorithm which needs further development.

Acknowledgments

This research was supported in part by NIH grants R01 EB013248, R01 EB008015, R01 LM010033, R01 NS079788, U01 NS082320, R42 MH086984, P30 HD018655, and by a research grant from Boston Children's Hospital Translational Research Program.

REFERENCES

1. Artaechevarria X, Munoz-Barrutia A. "Combination strategies in multi-atlas image segmentation: Application to brain MR data,". Medical Imaging, IEEE Transactions on. 2009; vol. 28(no. 8): 1266–1277.
2. 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. 2010; (no. 99):1.
3. 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. 2010; vol. 49(no. 3):2352–2365. [PubMed: 19857578]
4. 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
5. Akhondi-Asl A, Jafari-Khouzani K, Elisevich K, Soltanian-Zadeh H. "Hippocampal volumetry for lateralization of temporal lobe epilepsy: Automated versus manual methods,". Neuroimage. 2011; vol. 54:S218–S226. [PubMed: 20353827]
6. Khan A, Cherbuin N, Wen W, Anstey K, Sachdev P, Beg M. "Optimal weights for local multi-atlas fusion using supervised learning and dynamic information(SuperDyn): Validation on hippocampus segmentation,". NeuroImage. 2011; vol. 56(no. 1):126–139. [PubMed: 21296166]
7. Commowick O, Akhondi-Asl A, Warfield SK. "Estimating a reference standard segmentation with spatially varying performance parameters: Local MAP STAPLE,". Medical Imaging, IEEE Transactions on. 2012; vol. 31(no. 8):1593–1606.

8. Heckemann R, Hajnal J, Aljabar P, Rueckert D, Hammers A. “Automatic anatomical brain mri segmentation combining label propagation and decision fusion,”. NeuroImage. 2006; vol. 33(no. 1): 115–126. [PubMed: 16860573]
9. Warfield SK, Zou KH, Wells WM. “Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation,”. Medical Imaging, IEEE Transactions on. 2004; vol. 23(no. 7):903–921.
10. Asman A, Landman B. “Robust statistical label fusion through consensus level, labeler accuracy, and truth estimation (COLLATE),”. IEEE Transactions on Medical Imaging. 2011; vol. 30(no. 10):1779–1794. [PubMed: 21536519]
11. Landman B, Asman A, Scoggins A, Bogovic J, Xing F, Prince J. “Robust statistical fusion of image labels,”. IEEE Transactions on Medical Imaging. 2011; vol. 31(no. 2):512–522. [PubMed: 22010145]
12. Commowick O, Warfield SK. “Incorporating priors on expert performance parameters for segmentation validation and label fusion: a maximum a posteriori staple,” in. Medical Image Computing and Computer-Assisted Intervention(MICCAI’10), ser. LNCS. Sep.2010 vol. 6363:25–32.
13. Commowick O, Warfield SK. “A continuous STAPLE for scalar, vector and tensor images: An application to DTI analysis,”. IEEE Transactions on Medical Imaging. Jun.2009 vol. 28(no. 6): 838–846. [PubMed: 19272988]
14. Langerak TR, van der Heide UA, Kotte AN, Viergever MA, van Vulpen M, Pluim JP. “Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE),”. Medical Imaging, IEEE Transactions on. 2010; vol. 29(no. 12):2000–2008.
15. Asman A, Landman B. “Formulating spatially varying performance in the statistical fusion framework,”. Medical Imaging, IEEE Transactions on. 2012; vol. 31(no. 6):1326–1336.
16. 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. 2010; vol. 14(no. 1):39. [PubMed: 19897403]
17. Wang H, Suh JW, Das S, Pluta J, Altinay M, Yushkevich P. “Regression-based label fusion for multi-atlas segmentation,”. Computer Vision and Pattern Recognition (CVPR), *2011 IEEE Conference on*. IEEE. 2011:1113–1120.
18. Cardoso MJ, Leung K, Modat M, Keihaninejad S, Cash D, Barnes J, Fox NC, Ourselin S. “STEPS: Similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation,”. Medical Image Analysis. 2013
19. Asman A, Landman B. “Non-local statistical label fusion for multi-atlas segmentation,”. Medical Image Analysis. 2012
20. Weisenfeld NI, Warfield SK. “SoftSTAPLE: Truth and performance-level estimation from probabilistic segmentations,”. Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on IEEE. 2011:441–446.
21. Brouwer L. “Über abbildung von mannigfaltigkeiten,”. Mathematische Annalen. 1911; vol. 71(no. 1):97–115.
22. Kapur T. “Model based three dimensional medical image segmentation,”. Ph.D. dissertation, Massachusetts Institute of Technology. 1999
23. Celeux G, Forbes F, Peyrard N. “Em procedures using mean field-like approximations for markov model-based image segmentation,”. Pattern recognition. 2003; vol. 36(no. 1):131–144.
24. Rohlfing T, Maurer C. “Shape-based averaging,”. Image Processing, IEEE Transactions on. 2007; vol. 16(no. 1):153–161.
25. Wang H, Suh J, Das S, Pluta J, Altinay M, Yushkevich P. “Multi-atlas segmentation with joint label fusion,”. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2013; vol. 35(no. 3):611–623.
26. Thompson P, Woods R, Mega M, Toga A. “Mathematical/computational challenges in creating deformable and probabilistic atlases of the human brain,”. Human brain mapping. 2000; vol. 9(no. 2):81–92. [PubMed: 10680765]

27. Roland P, Zilles K. “Brain atlases-a new research tool,”. Trends in neurosciences. 1994; vol. 17(no. 11):458–467. [PubMed: 7531886]
28. Zacharaki E, Hogea C, Shen D, Biros G, Davatzikos C. “Non-diffeomorphic registration of brain tumor images by simulating tissue loss and tumor growth,”. Neuroimage. 2009; vol. 46(no. 3): 762–774. [PubMed: 19408350]
29. Durrleman S, Pennec X, Trouvé A, Thompson P, Ayache N. “Inferring brain variability from diffeomorphic deformations of currents: an integrative approach,”. Medical image analysis. 2008; vol. 12(no. 5):626. [PubMed: 18658005]
30. Thompson P, Schwartz C, Lin R, Khan A, Toga A. “Three-dimensional statistical analysis of sulcal variability in the human brain,”. The Journal of Neuroscience. 1996; vol. 16(no. 13):4261–4274. [PubMed: 8753887]
31. Blezek D, Miller J. “Atlas stratification,”. Medical Image Analysis. 2007; vol. 11(no. 5):443–457. [PubMed: 17765003]
32. 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. 2009; vol. 46(no. 3):726–738. [PubMed: 19245840]
33. Gerber S, Tasdizen T, Joshi S, Whitaker R. “On the manifold structure of the space of brain images,”. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009*, pp. 2009:305–312.
34. Commowick O, Warfield SK, Malandain G. “Using frankenstein’s creature paradigm to build a patient specific atlas,”. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009. 2009:993–1000.
35. Caviness VS, Filipek PA, Kennedy DN. “Magnetic resonance technology in human brain science: blueprint for a program based upon morphometry,”. Brain and Development. 1989; vol. 11(no. 1): 1–13. [PubMed: 2646959]
36. Kennedy DN, Filipek PA, Caviness VS Jr. “Anatomic segmentation and volumetric calculations in nuclear magnetic resonance imaging,”. Medical Imaging, IEEE Transactions on. 1989; vol. 8(no. 1):1–7.
37. Rohlfing T, Russakoff DB, Maurer CR Jr. “Expectation maximization strategies for multi-atlas multi-label segmentation,”. Information Processing in Medical Imaging. 2003; vol. 2732:210–221. [PubMed: 15344459]
38. 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. 2010; vol. 49(no. 3):2457–2466. [PubMed: 19818860]

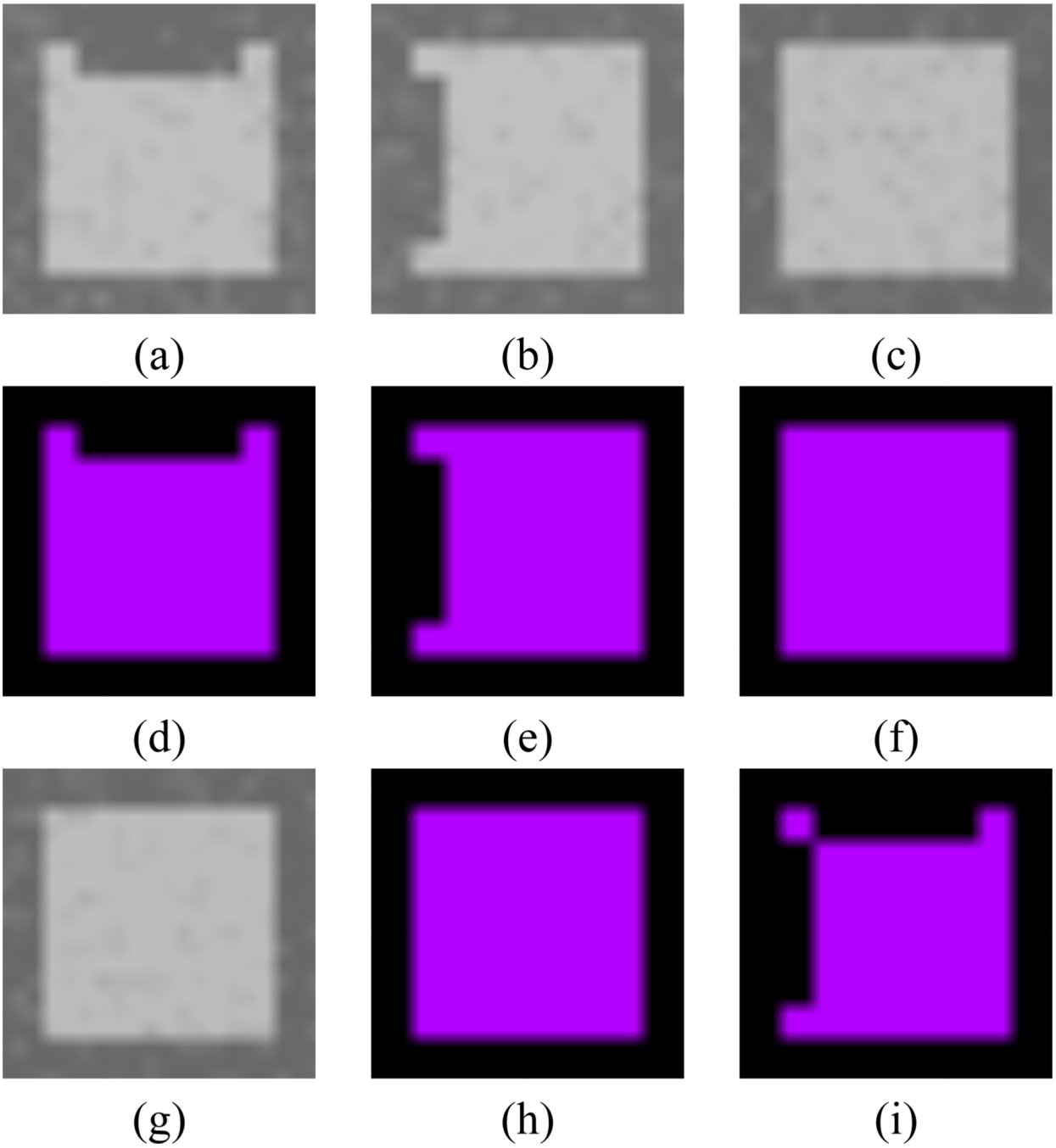

Fig. 1. Illustration of Synthetic Data Segmentation Results

Comparison of segmentation results generated by PSTAPLE (h), and STAPLE (i) using three samples from (a,d), three samples from (b,e), and two samples from (c,f). The target image is shown in (g) and the corresponding segmentation is shown in (h). It can be seen that using intensity information PSTAPLE has perfect segmentation accuracy. However, STAPLE could not find the correct answer because there is no way to distinguish between the template segmentations.

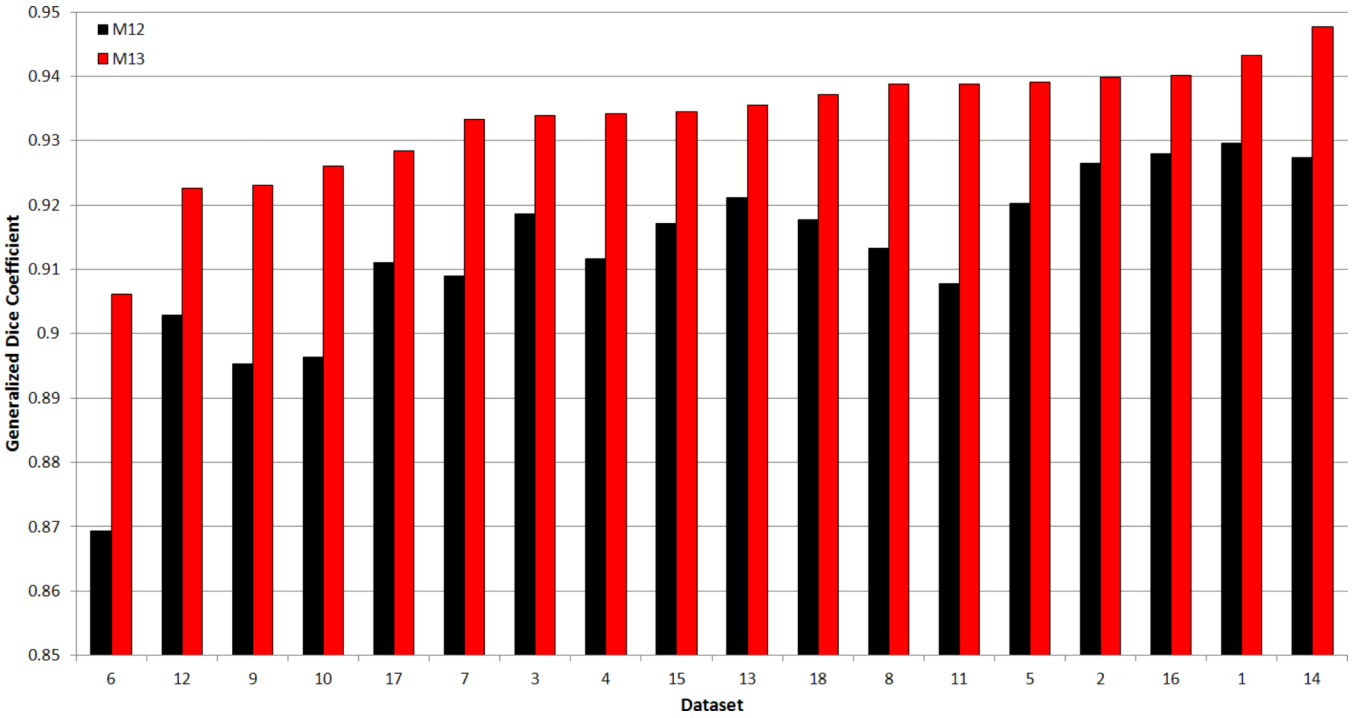

Fig. 2. Quantitative Comparison of the IBSR Multi-Atlas Segmentation Results

Comparison of generalized Dice coefficient of M_{12}: Local MAP STAPLE and M_{13}: Local MAP PSTAPLE for 34 structures in 18 IBSR datasets. It indicates that Local MAP PSTAPLE is superior to the Local STAPLE. The horizontal axis represents each subject.

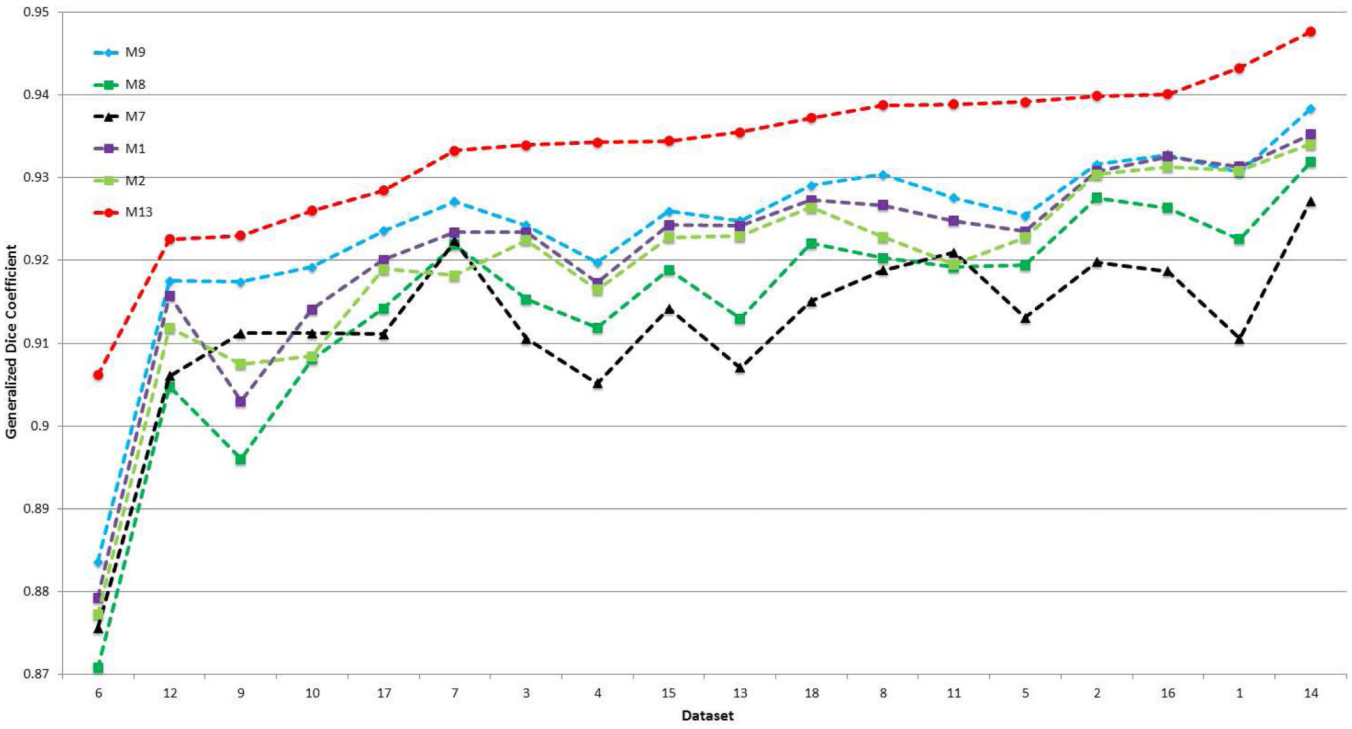

Fig. 3. Quantitative Comparison of the IBSR Multi-Atlas Segmentation Results

Comparison of generalized Dice coefficient of M_{13}: Local MAP PSTAPLE, M_1: method of Artaechevarria et al. [1], M_2: method of Sabuncu et al. [2], M_7: Multi-STEPS [18], M_8: NLS [19], and M_9: PICSL-MALF [25] for 34 structures in 18 IBSR datasets. It indicates that Local MAP PSTAPLE is superior to the other methods. The horizontal axis represents each subject.

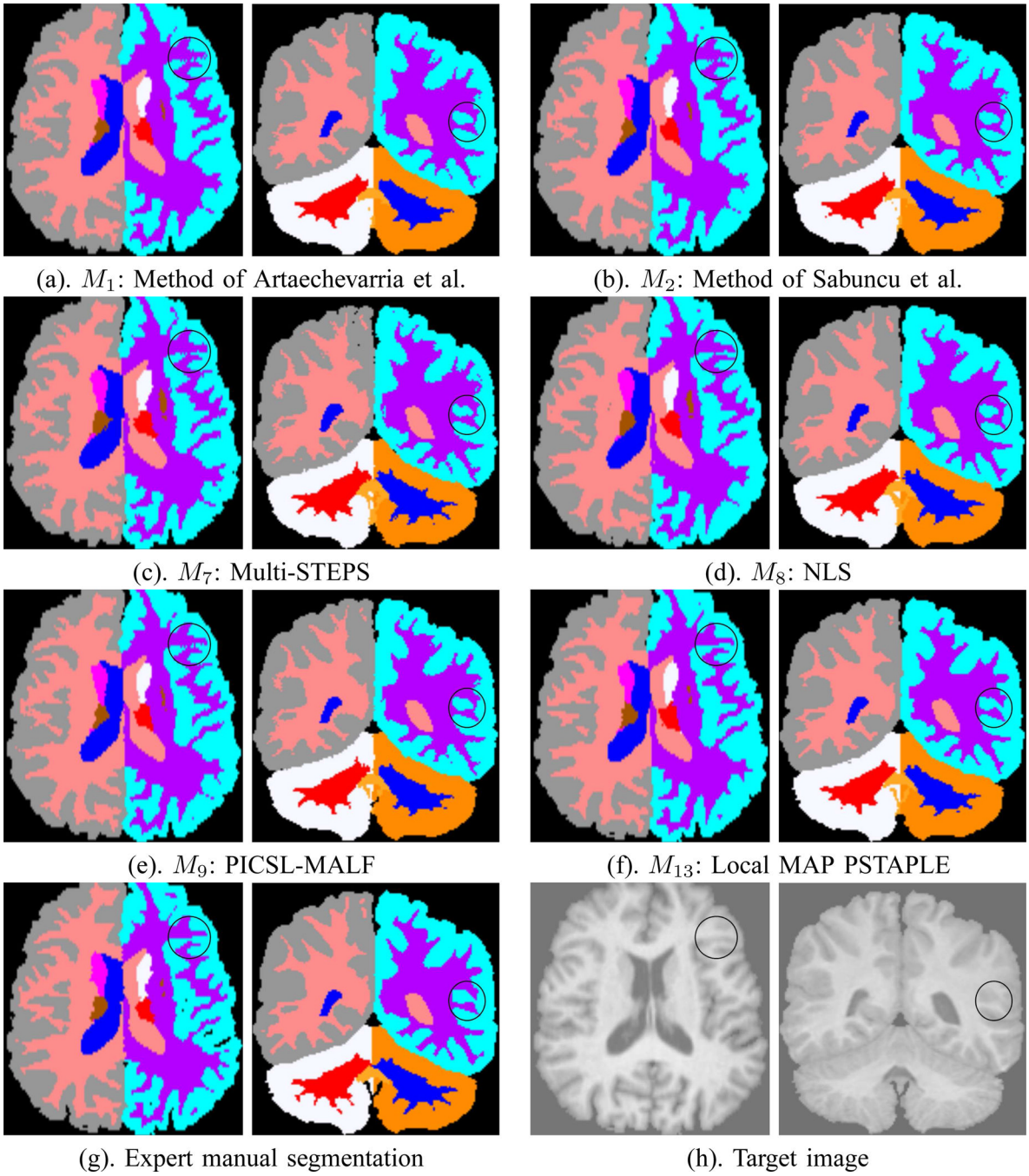

Fig. 4. Illustration of IBSR Multi-Atlas Segmentation Results

Comparison of segmentation results generated by M_{13}: Local MAP PSTAPLE (f), $M1$: method of [1] (a), M_2: method of [2] (b), M_7: Multi-STEPS [18] (c), M_8: NLS [19] (d), M_9: PICSL-MALF [25] (e), and expert manual segmentation (g) in an axial image of a representative IBSR dataset (h). Circles show the regions that Local MAP PSTAPLE clearly outperforms other methods.

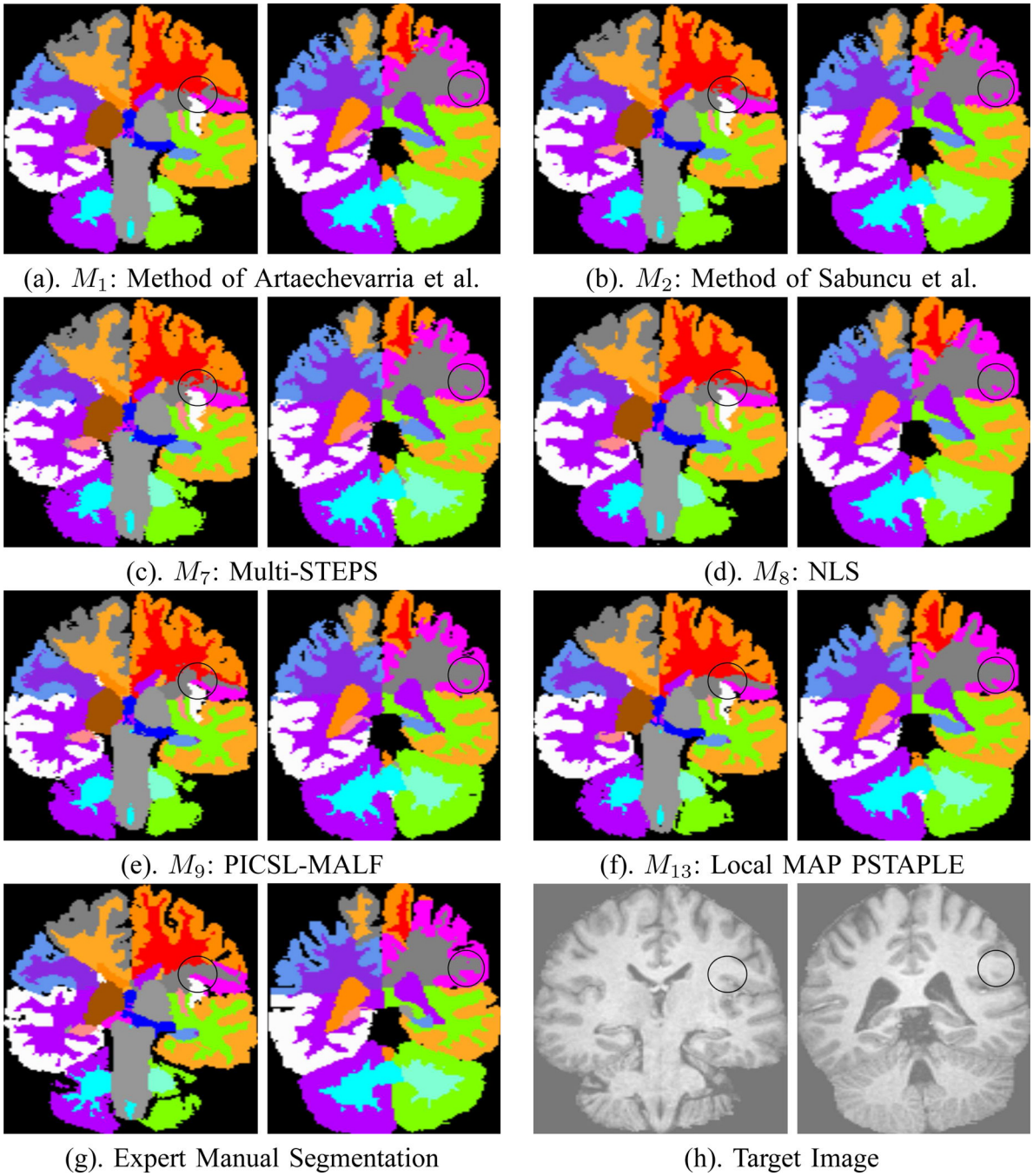

Fig. 5. Illustration of NMM Multi-Atlas Segmentation Results

Comparison of segmentation results generated by M_{13}: Local MAP PSTAPLE (f), M_1: method of [1] (a), M_2: method of [2] (b), M_7: Multi-STEPS [18] (c), M_8: NLS [19] (d), M_9: PICSL-MALF [25] (e) , and expert manual segmentation (g) in coronal images of representative NMM datasets (h). Circles show the regions that Local MAP PSTAPLE clearly outperforms other methods.

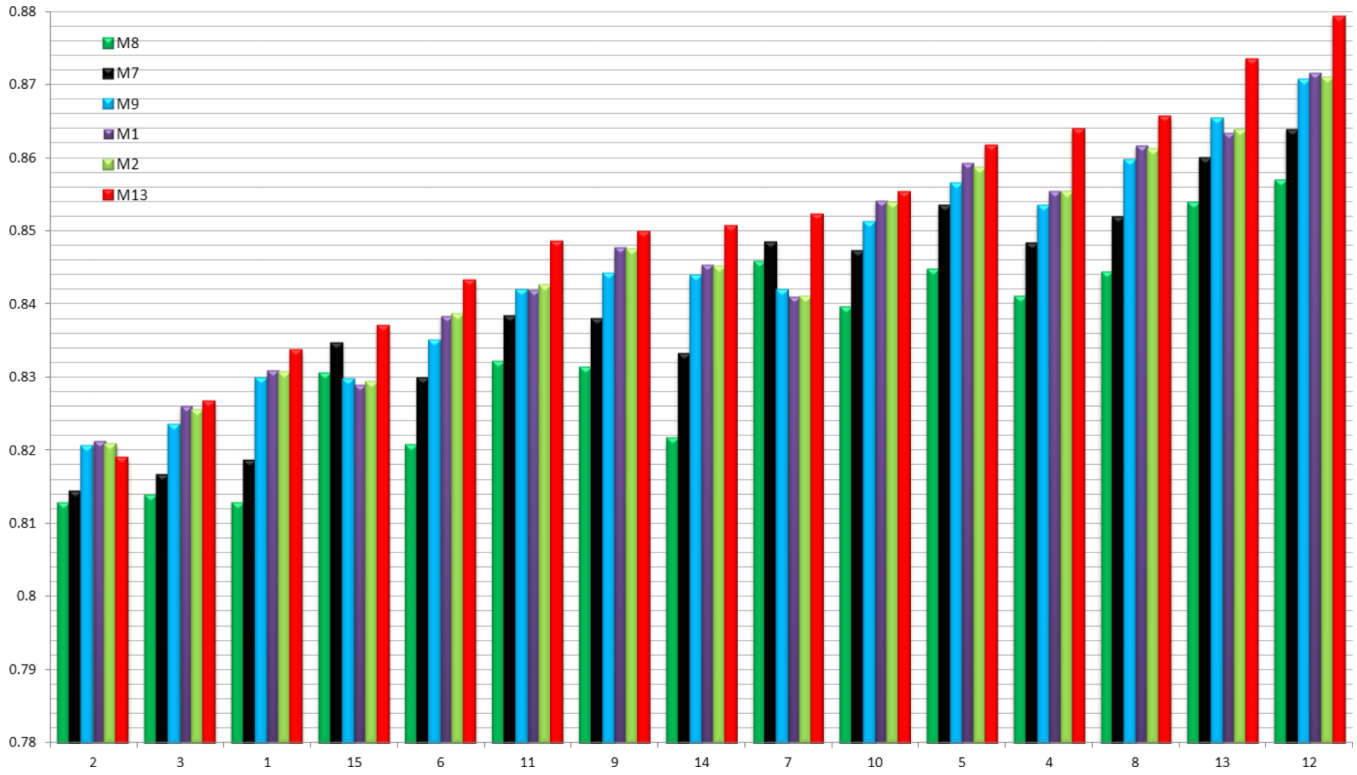

Fig. 6. Quantitative Comparison of the NMM Multi-Atlas Segmentation Results

Comparison of generalized Dice coefficient of M_{13}: Local MAP PSTAPLE and methods of M_1 [1], M_2 [2], M_7 [18], M_8 [19], and M_9 25 for 53 structures in 15 NMM datasets. It indicates that Local MAP PSTAPLE is superior to the other methods. The horizontal axis represents each subject.

TABLE I

Comparison of Dice coefficients for the proposed methods. Results of the segmentation of 16 × 2 + 1 = 33 sample structures and average of the 128 brain structures of the IBSR dataset for M_0:Majority Voting, M_1: Artaechevarria et al. method [1], M_2: Sabuncu et al. algorithm [2], M_3: SIMPLE [14], M_4: STAPLER [11],M_5 : COLLATE [10], M_6: STAPLE [9],M_7: Multi-STEPS [18], M_8: NLS [19],M_9: PICSL-MALF [25],M_{10}: Probabilistic Voting, M_{11}: PSTAPLE

Structure	M_0	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}
FP	0.845	0.850	0.848	0.841	0.808	0.846	0.847	0.849	0.850	0.834	0.861	**0.862**
F2	0.766	0.779	0.777	0.746	0.725	0.774	0.777	0.781	0.782	0.780	0.780	**0.787**
PRG	0.731	0.746	0.742	0.708	0.698	0.750	0.751	0.751	0.751	0.749	0.747	**0.759**
TP	0.835	0.840	0.839	0.818	0.791	0.834	0.837	0.841	0.837	0.845	0.851	**0.851**
POG	0.633	0.723	0.722	0.684	0.671	0.737	0.737	0.737	0.741	0.729	0.653	**0.747**
SGp	0.507	0.525	0.529	0.518	0.525	0.587	0.593	0.583	0.595	0.535	0.541	**0.602**
AG	0.216	0.254	0.253	0.343	0.419	0.382	0.434	0.421	0.451	0.279	0.233	**0.454**
Ols	0.766	0.780	0.779	0.749	0.727	0.773	0.776	0.782	0.782	0.781	0.793	**0.797**
Oli	0.654	0.674	0.673	0.632	0.627	0.684	0.687	0.689	0.697	0.683	0.679	**0.703**
PHp	0.646	0.658	0.656	0.623	0.607	0.665	0.667	0.673	0.675	0.668	0.651	**0.679**
OP	0.738	0.744	0.744	0.721	0.694	0.743	0.745	0.746	0.752	0.727	0.770	**0.773**
Cerebral White Matter	0.896	0.909	0.906	0.894	0.889	0.899	0.898	0.905	0.904	0.912	**0.918**	**0.918**
Inf Lat Vent	0.420	0.516	0.475	0.399	0.404	0.525	0.523	0.562	0.523	0.567	0.588	**0.639**
Cerebellum White Matter	0.876	0.880	0.878	0.874	0.842	0.875	0.876	0.875	0.875	0.882	0.887	**0.888**
Cerebellum Cortex	0.946	0.948	0.947	0.946	0.933	0.945	0.946	0.947	0.948	0.951	0.958	**0.958**
4th Ventricle	0.823	0.833	0.826	0.821	0.760	0.826	0.827	0.831	0.827	0.836	**0.843**	**0.843**
Caudate	0.852	0.867	0.858	0.857	0.823	0.860	0.858	0.869	0.870	**0.871**	0.858	0.868
Average of all 128 Structures	0.670	0.687	0.690	0.666	0.651	0.697	0.708	0.710	0.709	0.701	0.683	**0.717**