Automatic Optimization of Segmentation Algorithms Through Simultaneous Truth and Performance Level Estimation (STAPLE)

Mahnaz Maddah, Kelly H. Zou, William M. Wells, Ron Kikinis, and Simon K. Warfield

Computational Radiology Laboratory, Surgical Planning Laboratory, Brigham and Women's Hospital, Harvard Medical School, Boston MA 02115, USA.
{mmaddah, zou, sw, kikinis, warfield}@bwh.harvard.edu
http://spl.bwh.harvard.edu

Abstract. The performance of automatic segmentation algorithms often depends critically upon a number of parameters intrinsic to the algorithm. Appropriate setting of these parameters is a pre-requisite for successful segmentation, and yet may be difficult for users to achieve. We propose here a novel algorithm for the automatic selection of optimal parameters for medical image segmentation. Our algorithm makes use of STAPLE (Simultaneous Truth and Performance Level Estimation), a previously described and validated algorithm for automatically identifying a reference standard by which to assess segmentation generators. We execute a set of independent automated segmentation algorithms with initial parameter settings, on a set of images from any clinical application under consideration, estimate a reference standard from the segmentation results using STAPLE, and then identify the parameter settings for each algorithm that maximizes the quality of the segmentation generator result with respect to the reference standard. The process of estimating a reference standard and estimating the optimal parameter settings is iterated to convergence.

1 Introduction

The analysis of medical images is a critical process, enabling applications ranging from fundamental neuroscience, to objective evaluation of interventions and drug treatments, to monitoring, navigation and assessment of image guided therapy. Segmentation is the key process by which raw image acquisitions are interpreted. Interactive segmentation is fraught with intra-rater and inter-rater variability which limits its accuracy, while also being costly and time-consuming. Automatic segmentation holds out the potential of dramatically increased precision, and reduction in time and expense. However, the performance of automatic segmentation algorithms often depends critically upon a number of parameters intrinsic to the algorithm. Such parameters may control assumptions regarding tissue intensity characteristics, spatial homogeneity constraints, boundary smoothness or curvature characteristics or other prior information. Appropriate setting of these parameters by users is often a prerequisite for successful segmentation, and yet may be difficult for users to achieve

C. Barillot, D.R. Haynor, and P. Hellier (Eds.): MICCAI 2004, LNCS 3216, pp. 274–282, 2004.

due to the potentially nonlinear effects and interactions between different parameter settings which are challenging for human operators to optimize over. Furthermore, the selection of parameters based on a synthetic phantom may not be appropriate for clinical applications since the normal and pathological appearance of subjects from any particular clinical population may be quite different from that readily captured in a phantom.

To overcome these problems, we present here an approach to estimate the true segmentation from several automatic or semi-automatic segmentation algorithms and optimize their free parameters for a category of medical images. The ground-truth estimation is done by an Expectation-Maximization algorithm, called STAPLE (Simultaneous Truth and Performance Level Estimation), presented in [1]. In that work, a collection of expert-segmented images was given to STAPLE and a probabilistic estimate of the true segmentation was computed along with the performance level measurement of each expert. The idea here is to employ this method in order to obtain the optimal values for the parameters of different automated segmentation algorithms and to evaluate their performance compared with the estimated true segmentation. This is significantly important in medical application such as neuroscience and surgical planning, since the users often face difficulty to find good parameter settings and yet their results are utilized for the research studies and disease therapies.

The paper is organized as follows: In Section 2, we outline our evaluation and parameter optimization methodology. The ground-truth estimation, the optimization algorithm, and the assessment metric used in this study are described in this section. The optimization and evaluation of the four algorithms for brain tissue segmentation as well as the obtained results are presented in Section 3, and finally the conclusion and further work are brought in Section 4.

2 Methods

The proposed method is an iterative process with two main stages: ground-truth estimation and parameter setting which can be performed on a specific case or a collection of subjects:

a) Optimization on One Subject:

To find the best parameters for each subject, we first need to estimate a ground truth and performance level on a training dataset segmented using the algorithms under study. In this stage, the segmented images from different algorithms are given to STAPLE which computes simultaneously a probabilistic estimate of the true segmentation and a measure of the performance level represented by each segmentation algorithm. Then, we optimize the performance of each algorithm with respect to the estimated ground truth. Given a set of algorithms with optimized parameters, we recompute the ground truth and re-optimize the parameters until the ground truth estimate converges. At the end of this stage we have the optimized segmentation algo-

rithms, an estimated true segmentation, and the levels of performance for each algorithm on one experimental data.

b) Optimization on a Collection of Subjects:

The goal of this step would be to expand the capability of the approach to optimize parameters of each algorithm across N training subjects, for $N > 1$.This is achieved by changing the quality measure to be the mean of the performance level across all cases. Each optimization step is performed based on the impact of the parameter across the N cases rather than just one case. This would give us optimized parameters for a set of subjects, finding a tradeoff in settings across all of the subjects. It might do worse than possible if we optimized just for one case, but on average leads to a better result over all the subjects.

In fact, our approach is technically an instance of a generalized expectation maximization algorithm, where we have extra parameters (the segmentation algorithm parameters) for which no closed form maximization exists and so an approximate local maximization strategy is used.

2.1 STAPLE

STAPLE takes a collection of segmentations of an image, and constructs a probabilistic estimate of the true segmentation and a measure of the performance level of each segmentation generator [1]. This algorithm is an instance of the Expectation-Maximization (EM) in which the segmentation decision at each voxel is directly observable, the hidden true segmentation is a binary variable for each voxel, and the performance level, achieved by each segmentation method is represented by sensitivity and specificity parameters [1]. At each EM iteration, first, the hidden true segmentation variables are replaced with their conditional probabilities and are estimated given the input segmentation and a previous estimate of the performance level. In the second step, the performance parameters are updated. This process is iterated until the convergence is reached. STAPLE is also capable of considering several types of spatial constrains, including a statistical atlas of prior probabilities for the distribution of the structure of interest which we make use of it in our approach.

2.2 Performance Metric

The "ideal" performance point represents true positive rate (TP) of 1 and false positive rate (FP) of 0. With the TP and FP obtained by comparing each segmentation with the estimated ground truth, a performance metric can be defined based on the weighted Euclidean distance from this point:

$$Performance = 1 - \sqrt{w^2 (1 - TP)^2 + (1 - w)^2 FP^2} \tag{1}$$

The weighting factor, w, is set equal to the foreground prevalence. When the foreground occupies a small portion of the image (small w), TP is very sensitive to the

number of foreground errors, while *FP* is relatively insensitive to the errors in the background. Weighting the errors makes the overall error almost equally sensitive to the errors in the foreground and background.

2.3 Optimization Method

Numerous optimization algorithms exist in the literature and each might be invoked in our application provided that it is able to find the global optimum and can be applied to discrete optimization problems.

Here, we make use of simultaneous perturbation stochastic approximation (SPSA) method [2]. It requires only two objective function measurements per iteration regardless of the dimension of the optimization problem. These measurements are made by simultaneously varying in a proper random fashion of all the parameters. The random shift of parameters is controlled by a set of SPSA algorithm coefficients which must be set for each algorithm under-optimization. The error function here, is 1–*Performance* computed for each segmentation from equation (1). Note that the behavior of the error in terms of the parameter should be known in order to correctly set the acceptable range of parameters.

3 Experiments and Results

We have applied our method to the problem of tissue segmentation of human's brain, which has received continues attention in medical image analysis, focusing on white mat classification.

3.1 Human Brain Tissue Segmentation Algorithms

We considered four algorithms including two research segmentation algorithms and two well-known packages which are briefly described in the following:

SPM – SPM uses a modified maximum likelihood "mixture model" algorithm for its segmentation, in which each voxel is assigned a probability of belonging to each of the given clusters [3]. Assuming a normal distribution for the intensity of the voxels belonging to each cluster, the distribution parameters are computed. These parameters are then combined with a given priority map to update the membership probabilities. The smoothing parameter applied to affine registration is considered as the parameter to be optmimally set.

FSL – The segmentation algorithm in this package is based on a hidden Markov random field (HMRF) model [4]. It starts with an initial estimation step to obtain initial tissue parameters and classification, followed by a three-step EM process which updates the class labels, tissue parameters and bias field iteratively. During the teratn MRF-MAP (maximum a posteriori) approach is used to estimate class labels, mp is applied to estimate the bias field, and the tissue parameters are estimated by

maximum likelihood (ML). We used the MRF neighborhood beta value as the controlling parameter for optimization.

***k*-NN Classification –** The *k*-Nearest Neighbor (*k*-NN) classification rule is a technique for nonparametric supervised pattern classification. Given a training data set consisting of N prototype patterns and the corresponding correct classification of each prototype into one of *C* classes, a pattern of unknown class, is classified as class *C* if most of the closest prototype patterns are from class *C*. The algorithm we used is a fast implementation of this basic idea[5]. The number of training data patterns, their position and K, the number of nearest neighbors to consider, are the controlling factors of this method which the latter, K is considered here as the parameter to be optimized.

EM – This method is an adaptive segmentation, which uses an Expectation Maximization (EM) algorithm [6]. It simultaneously labels and estimates the intensity inhmogeneities artifacts in the image. It uses an iterated moving average low-pass filter in bias field estimation which its width is considered as the parameter for our study. We set the number of EM iteration steps to 10 according to [6].

3.2 Training and Test Data

We applied the approach on five sets of T1-weighted MR brain images with resolution of 0.9375×0.9375×1.3 mm^3, each consists of 124 slices. The images were first filtered by a multi-directional flux-diffusion filter, implemented in 3D Slicer [7]. Next, the non-brain tissues are removed from the images by brain extraction tool (BET) in the FSL package which uses a deformable model to fit the brain's surface [8]. The average of the semi-automatic segmentation of 82 cases was used as the atlas for STAPLE.

3.3 Optimization Results

In Table I, the optimization results for each of the four algorithms are shown. In the first step (iteration 0), each algorithm has been run with the default value of its parameter. With the estimated ground-truth in Iteration *i*, each algorithm goes through the optimization loop and the obtained optimal values are set for the next ground-truth estimation (iteration *i*+1). The iteration stops when the parameters converge. Once the optimized parameters are found for one case randomly selected from the pool of datasets, we use these values as the initial parameter setting for optimizing on a collection of datasets. The *k*-NN algorithm is an exception to this, as discussed later. The sensitivity (p) and specificity (q) obtained from STAPLE, are given in Table I as the performance measures.

EM – As seen in Table I, this algorithm gains a high performance score from STAPLE. The error rate versus its parameter, *i.e*. the width of the low-pass filter for bias field estimation in pixels, is shown in Fig. 1(a) for one of the cases under study. The minimum error corresponds to relatively large value of 90 which reflects the low RF coil artifact in the image. A similar trend was observed for other cases.

Table 1. The sensitivity (p) and specificity (q) of each segmentation algorithms are given for the iterations of optimization process. In Iteration 0, STAPLE runs given the four segmentations extracted by each method with the corresponding initial parameters. The new parameters are obtained with respect to the estimated ground-truth in iteration 0. In Iteration 1, the STAPLE runs with the obtained parameters and a new ground-truth and (p,q) pairs are estimated. No changes in parameters occurs in the next optimization, so we stop. Note that for optimization on five cases, the mean and the standard deviation of (p,q)'s over all cases are given in the Table.

		Optimization on one case			Optimization on five cases	
Alg.		Iter. 0	Iter. 1		Iter. 0	Iter. 1
SPM	Window Len.	8	12		12	10
	p	0.6866	0.6843		0.5912±0.1085	0.6014±0.0999
	q	0.9996	0.9995		0.9943±0.0073	0.9966±0.0062
FSL	MRF β	0.3	0.2		0.2	0.1
	p	0.9017	0.9135		0.8149±0.1664	0.8498±0.1892
	q	0.9990	0.9983		0.9960±0.0068	0.9955±0.0063
k-NN	K	9	6		15	12
	p	0.9983	0.9554		0.9497±0.0480	0.9574±0.0482
	q	0.9937	0.9988		0.9986±0.0020	0.9987±0.0020
EM	Filter Width	31	90		90	90
	p	0.9000	0.9899		0.9890±0.0164	0.9857±0.0191
	q	1.0000	0.9990		0.9955±0.0043	0.9951±0.0052

***k*-NN –** This algorithm also gets a good performance score. However the resulting optimum K, number of neighbors to be considered, is smaller than its default value 9 which is inconsistent with the fact that the error rate in k-NN method is inversely proportional to K. This can be investigated by looking at Fig. 1(b) in which the error vs. the parameter K is illustrated. Although, the graph shows a minimum at K = 6 (which might be due to excluding specific prototypes), the error is very sensitive to the choice of K in that region. A better parameter setting is to set K>12 to avoid the transition region (Note that the upper limits for K is the smallest number of prototypes, selected by the user for each tissue class). In order to prevent the optimization algorithm from stopping in such unwanted minima, one can define the optimization goal as a combination of the error and the error sensitivity to the parameter.

FSL – The optimum value for MRF neighborhood beta is obtained to be 0.2 when considering one subject and 0.1 for our collection of subjects. As seen in Fig. 1(c), the error rate is a well-behaved function of the parameter, however contrary to EM and k-NN methods, the error rate increases as the optimization algorithm proceeds. This is due to the fact that the estimated ground-truth converges to EM and k-NN results, and therefore the difference between the segmentation by FSL and the estimated ground-truth increases more and more.

SPM – This algorithm underestimates the white matter tissue compared to the prevailing algorithms (EM and k-NN), and thus, a low performance level is assigned to it by STAPLE. Furthermore, the selected parameter is not able to push the segmentation results to the estimated ground-truth, though the improvement is apparent in Fig.2 (e). The algorithm gets even lower performance level in the second run, as can be observed in Fig. 1(d).

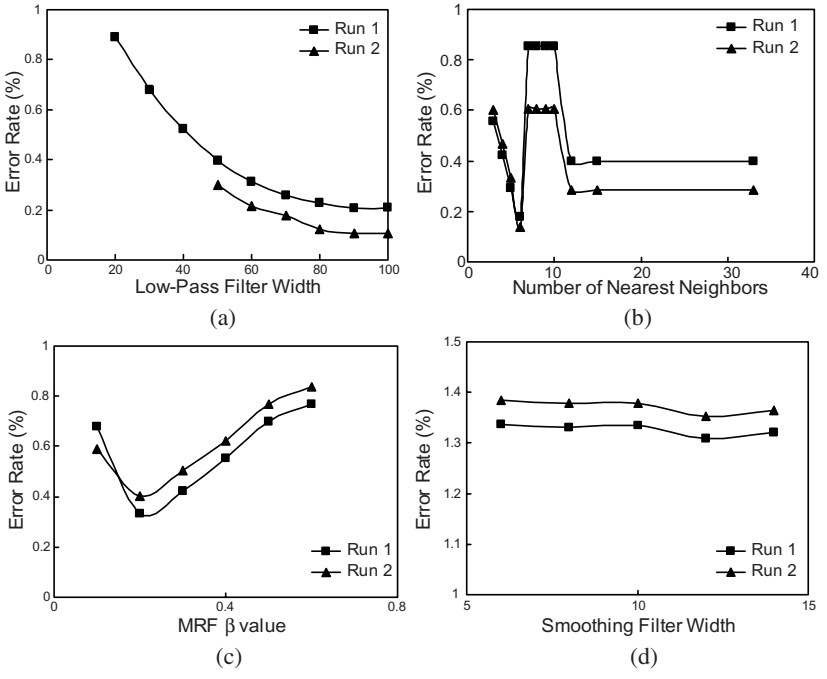

Fig. 1. The error rates versus the parameter for (a) EM (b) *k*-NN (c) FSL and (d) SPM algorithms after the first and second ground-truth estimations. The overall error rate obtained for EM and *k*-NN algorithms decreases as the estimated ground-truth approaches the results of these two algorithms.

The effect of parameter adjustment can be observed in Fig.2, where the segmentations with the default and optimized parameters are illustrated.

4 Conclusion and Further Work

In this paper, we presented a novel approach for the automatic selection of optimal parameters for segmentation. This algorithm is an iterative process with two main stages: ground-truth estimation and parameter setting. Two widely-used packages, SPM and FSL, and two research algorithms, k-NN and EM were considered for the optmization. We applied these algorithms on a set of 3D MR images of brain to segment the white matter tissue. The optimal parameters were obtained first for a single case and then for a collection of five cases. The process of estimating a ground-truth and estimating the optimal parameter settings converges in a few iterations. The proposed approach was shown to provide improved segnmentation results compared with algorithms with default parameter setting.

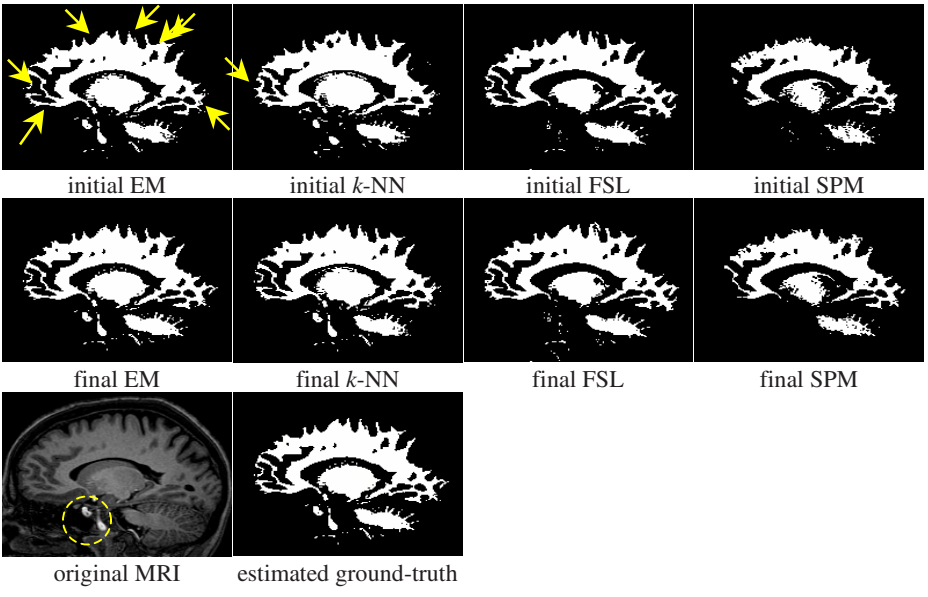

Fig. 2. The optimized segmentations are improved as compared to the initial segmentations. White matter segmentations obtained from different algorithms before and after optimization along with the original MR image and the final estimated ground truth. Some of the errors in the initial segmentation corrected by the proposed algorithm are highlighted by arrows.

Further work is under way to apply this approach to more medical images, to segmentation of other structures. Removing the artifacts in the images such as that highlighted in Fig. 2 with a circle, is another important step to improve the final ground-truth estimation.

Acknowledgement. This investigation was supported by a research grant from the Whitaker Foundation, and by NIH grants R21 MH67054, R01 LM007861, P41 RR13218, P01 CA67165.

References

1. Warfield, S. K., Zou, K. H., Wells, W. M.: Validation of Image Segmentation and Expert Quality with an Expectation-Maximization Algorithm. MICCAI 2002: 5th International Conference, Tokyo, Japan, September 25-28, Proceedings, Part I. (2002) 298 – 306.
2. Spall, J. C.: Stochastic Approximation and Simulated Annealing. Encyclopedia of Electrical and Electronics Engineering, Wiley, New York, Vol. 20, (1999) 529–542.
3. Ashburner, J.: Computational Neuroanatomy. PhD Thesis, University College London (2000). Available at http://www.fil.ion.ucl.ac.uk/spm.

4. Zhang, Y., Brady, M., and Smith S.: Segmentation of Brain MR Images Through a Hidden Markov Random Field Model and the Expectation-Maximization Algorithm. IEEE Trans. Medical Imaging, Vol. 20, No. 1, (2001) 45-57.
5. Warfield, S. K.: Fast *k*-NN Classification for Multichannel Image Data. Pattern Recognition Lett., Vol. 17, No. 7 (1996) 713-721.
6. Wells, W.M., Grimson, W.E.L., Kikinis, R., Jolesz, F.A.: Adaptive Segmentation of MRI data. IEEE Trans. Medical Imaging. Vol. 15, (1996) 429-442.
7. Krissian, K.: Flux-Based Anisotropic Diffusion: Application to Enhancement of 3D Angiogram. IEEE Trans. Medical Imaging, Vol. 22, No. 11 (2002)1440-1442.
8. Smith, S.: Fast Robust Automated Brain Extraction. Human Brain Mapping, Vol.17, No. 3, (2002) 143-155.