Combining Multiple Expert Annotations Using Semi-Supervised Learning And Graph Cuts For Crohn's Disease Segmentation

Dwarikanath Mahapatra[1,*], Peter J. Schüffler[1], Jeroen A.W. Tielbeek[2], Carl Puylaert[2], Jesica C. Makanyanga[3], Alex Menys[3], Rado Andriantsimiavona[4], Jaap Stoker[2], Stuart A. Taylor[3,5], Franciscus M. Vos[2,6], Joachim M. Buhmann[1]

[1]Department of Computer Science, ETH Zurich, Switzerland.
[2] Department of Radiology, Academic Medical Center, The Netherlands.
[3] Centre for Medical Imaging, University College London, United Kingdom.
[4] Biotronics3D, London, United Kingdom.
[5] University College London Hospitals, United Kingdom.
[6] Quantitative Imaging Group, Delft University of Technology, The Netherlands.
[*] *dwarikanath.mahapatra@inf.ethz.ch*

Abstract. We propose a graph cut (GC) based approach for combining annotations from multiple experts and segmenting Crohns disease (CD) tissues in magnetic resonance (MR) images. Random forest (RF) based semi supervised learning (SSL) predicts missing expert labels while a novel self consistency (SC) score quantifies the reliability of each expert label and also serves as the penalty cost in a second order Markov random field (MRF) cost function. The final consensus label is obtained by GC optimization. Experimental results on synthetic images and real CD patient data show our final segmentation to be more accurate than those obtained by competing methods. It also highlights the effectiveness of SC score in quantifying expert reliability and accuracy of SSL in predicting missing labels.

1 Introduction

Greater awareness about the seriousness of Crohn's disease (CD) within the medical imaging community has led to machine learning (ML) based analysis of magnetic resonance (MR) images to segment diseased regions [12, 8, 9] and predict disease severity [11]. Success of ML segmentation algorithms depend to a large extent on the accuracy of expert annotations. It is a common practice in medical image segmentation to obtain annotations from multiple experts, although combining them is not trivial since manual segmentations tend to be subjective, prone to inter-observer and intra-observer variability, and of varying accuracy. We propose to combine multiple expert annotations using semi supervised learning (SSL) and graph cuts (GC). The consensus annotation is used to design a ML approach for segmenting regions with CD activity.

One of the first methods to combine multiple annotations was STAPLE ([13]) which employed Expectation-maximization (EM) to find sensitivity and specificity values that maximize the data likelihood. MAP-STAPLE [6], used MRFs

to incorporate spatial constraints in STAPLE and generate a spatially consistent estimation of the ground truth. Commowick et al. in [5] adapt the STAPLE algorithm to determine spatially varying performance levels using sliding windows. The above algorithms fuse multiple labels independently from the original images, and hence do not assess their visual consistency. Raykar et al. [10] incorporate visual consistency of labels by simultaneous estimation of performance and learning. Chatelain et al. in [4] use Random forests (RF) to determine most coherent expert decisions based on the consistency of decisions with respect to the image features but do not account for missing annotations.

The following factors are important to obtain a consensus annotation of multiple experts: 1) predict missing annotations; 2) fuse annotations according to the reliability or consistency of experts; and 3) ensure spatial consistency of the final annotation. To achieve the above objectives we predict missing labels using SSL, quantify the reliability of each expert using a novel self-consistency (SC) score, and use Markov random fields (MRF) to impose spatial smoothness. SC is also used to define MRF penalty costs in the absence of true label information. The final (ground truth) annotation is obtained using GC optimization and is used to design a ML method to segment CD regions from unseen patient data.

Our work has two novelties: 1) a novel SC score to quantify the consistency and accuracy of each expert. It is calculated for each voxel from spatial feature distributions, and relates the annotations to image features. 2) missing expert labels are predicted using SSL that exploits the information from existing expert labels. Previous methods employed an iterative EM approach for this purpose while SSL predicts missing labels in one step. Graph cuts are used to determine the final labels because: a) no iterative approach is employed as in EM based approaches of [5, 6]; and b) globally optimum labels can be obtained thus reducing chances of getting stuck in local minima. We describe our method in Section 2, present our results in Section 3 and conclude with Section 4.

2 Method

2.1 Predicting Missing Labels

Let us consider a multi-supervised learning scenario with a training set $S = \{(x_n, y_n^1, \cdots, y_n^r)\}_{r=1}^{R}$ of samples x_n, and the corresponding labels y_n^r provided by R experts. Missing labels are commonly encountered when multiple experts annotate data. In previous approaches ([6, 5]) missing labels were predicted by combining Maximum A Posteriori (MAP) with iterative EM optimization to maximize the joint likelihood. We use semi-supervised RF classifiers (RF-SSL) to predict the missing labels by using knowledge from the given labels and image features. Unlike previous methods ([3]), a 'single shot' RF method for SSL without the need for iterative retraining was introduced in [7]. We use this SSL classifier as it is shown to outperform other approaches.

For labeled samples the information gain over data splits at each node of the RF is maximised and encourages separation of the labeled data [7, 2]. However

for SSL the objective function encourages separation of the labeled training data and simultaneously separates different high density regions. It is achieved via the following mixed information gain:

$$I_j = I_j^U + \alpha I_j^S \tag{1}$$

where $I_j^S = H(S_j) - \sum_{i \in \{L,R\}} \frac{|S_j^i|}{|S_j|} H(S_j^i)$ is the information gain from the labeled data; H is the entropy of training points, and S_j^L and S_j^R the subsets going to the left and right children of node j. I_j^U depends on both labeled and unlabeled data, and is defined using differential entropies over continuous parameters as

$$I_j^U = \log |\Lambda(S_j)| - \sum_{i \in \{L,R\}} \frac{|S_j^i|}{|S_j|} \log |\Lambda(S_j)| \tag{2}$$

Λ is the covariance matrix of the assumed multivariate distributions at each node. For further details we refer the reader to [7]. Thus the above cost function is able to combine the information gain from labeled and unlabeled data without the need for an iterative procedure.

To reduce computation time we select a region of interest (ROI) by taking the union of all expert annotations and determining its bounding box rectangle. The size of the rectangle is expanded by ± 20 pixels along rows and columns to give the final ROI. For each ROI pixel we calculate the mean and variance of intensity and 2D curvature values from a 15×15 neighborhood to give 4 features. Additionally, we extract spatial context features using the sampling template shown in Fig. 1 (a). The circle center is the current voxel and at each point corresponding to a red 'X' we calculate the mean intensity, and curvature values from a 3×3 window. The 'X's are located at distances of $3, 6, 9, 12$ pixels from the center, and the angle between consecutive rays is $45°$. 64 context features are obtained from the 32 points and the final vector has $(64 + 4 =)68$ values.

Each voxel has $r(\leq R)$ known labels and the unknown $R - r$ labels are predicted by SSL. The feature vectors of all samples (labeled and unlabeled) are inputted to the RF-SSL classifier which returns the missing labels. Note that although the same sample (hence feature vector) has multiple labels, RF-SSL treats it as another sample with similar feature values as other samples. The missing labels are predicted based on the split configuration (of decision trees in RFs) that leads to maximal global information gain. Hence the prediction of missing labels is not directly influenced by the other labels of the same sample but takes into account global label information [7].

2.2 Self Consistency of Experts

Self consistency of expert annotations is important to determine the reliability of each annotator. Region with similar labels are expected to have a consistent distribution of features. Figure 1 (b) shows an example annotation in which the annotated diseased boundary is shown in red. For any given point i (indicated

by the arrow) of the annotated region (labeled diseased or normal) we calculate the distribution of intensity values (normalized to be in $[0, 1]$) over a 15×15 neighborhood (yellow square). Then we compare the corresponding distributions of all similarly labeled points (diseased or normal) within a 35×35 neighborhood (points within the larger green square). The difference between two feature distributions is given by the χ^2 distance,

$$S(p, q) = \frac{1}{2} \sum_{k=1}^{K} \frac{[h_p(k) - h_q(k)]^2}{h_p(k) + h_q(k)}, \tag{3}$$

where p, q are two points under consideration, h_p, h_q are the respective normalized histograms and k denotes the k^{th} bin of the K-bin normalized histograms. $S \in [0, 1]$ is used as it gives a normalized distance measure with 0 indicating identical distributions. The self-consistency score for a point i (SC_i) is derived from the average S with respect to all similarly labeled points within the larger green square. It is defined as

$$\mathrm{SC}_i = 1 - \frac{1}{N_l} \sum_{n_l=1}^{N_l} S(i, n_l) \tag{4}$$

$S(i, n_l)$ denotes the χ^2 distance between i and all similarly labeled points n_l within the green square. N_l is the set of points within the 35×35 window with the same label as i.

An expert with high consistency will assign similar labels to regions having similar features. Thus by comparing the intensity distribution of a point i with other points of the same label, we get an estimate of its consistency. Boundary pixels of high consistency annotations have similar histograms, giving high SC. Non-boundary foreground pixels in the 35x35 window that are far from the boundary have different distributions and are fewer compared to foreground pixels on or near the boundary. Thus for boundary points neighboring pixels within the 35x35 window have similar features and give a reliable SC score. A point with high consistency has low value of S. Since we assign higher score for higher segmentation consistency the formulation as given in Eqn.4 is used.

We average $S(i, n_l)$ over points within a 35×35 neighborhood only, as feature consistency is higher in the local neighborhood. If we consider all the labeled points in the annotation, S would be biased towards the predominant label. It also reduces the computation time. Note that here we define the consistency score of each voxel i which is required for obtaining the ground truth segmentation. An overall consistency score for each annotation can be obtained by averaging the SC_i over all constituent pixels.

2.3 Obtaining the final labels

A second order MRF cost function is given by,

$$E(L) = \sum_{s \in P} D(L_s) + \lambda \sum_{(s,t) \in N_s} V(L_s, L_t), \tag{5}$$

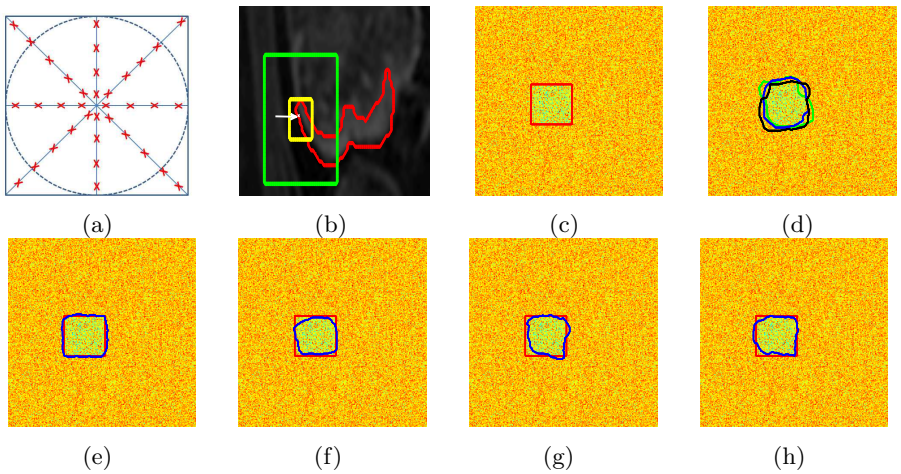

Fig. 1. (a) template for context feature extraction; (b) calculation of self consistency score; (c) synthetic image with ground truth segmentation in red; (d) synthetic image with simulated expert annotations; final segmentation obtained by (e) GC_{ME} (DM= 0.94); (f) STAPLE(DM= 0.90); (g) MAP-STAPLE(DM= 0.88); (h) Local MAP STAPLE(DM= 0.87).

where P denotes the set of pixels; N_s is the 8 neighbors of pixel s (or sample x); L_s is the label of s; t is the neighbor of s, and L is the set of labels for all s. $\lambda = 0.02$ determines the relative contribution of penalty cost (D) and smoothness cost (V). We have only 2 labels ($L_s = 1/0$ for object/background), although our method can also be applied to the multi-label scenario. The final labels are obtained by graph cut optimization [1].

The penalty cost for MRFs is normally calculated with respect to a reference model of each class (usually distribution of intensity values). The implicit assumption is that the annotator's labels are correct. However, we aim to determine the actual labels of each pixel and hence do not have access to true class distributions. To overcome this problem we use the consistency scores of experts to determine the penalty costs for a voxel. Each voxel has R labels (after predicting the missing labels). Say for voxel x the label y^r (of the rth expert) is 1, and the corresponding SC score is SC_x^r (Eqn.4). Since SC is higher for better agreement with labels, the corresponding penalty cost for $L_x = 1$ is

$$D(L_x = 1)^r = 1 - SC_x^r, \tag{6}$$

where L_x is the label of voxel x. The penalty cost for label 0 is

$$D(L_x = 0)^r = 1 - D(L_x = 1) = SC_x^r. \tag{7}$$

The final penalty costs for each L_x is the average of costs from each expert,

$$\begin{aligned} D(L_x = 1) &= \tfrac{1}{R} \textstyle\sum_{r=1}^{R} D(L_x = 1)^r, \\ D(L_x = 0) &= \tfrac{1}{R} \textstyle\sum_{r=1}^{R} D(L_x = 0)^r. \end{aligned} \tag{8}$$

Since iterative approaches may get stuck in local minima, GC optimization is appealing as it gives a global minima for binary labeled problems.

Smoothness Cost (V): V penalizes discontinuities amongst neighboring voxels and is a function of their intensity differences. V is given by

$$V(L_s, L_t) = \begin{Bmatrix} e^{-\frac{(I_s - I_t)^2}{2\sigma^2}} \cdot \frac{1}{\|s-t\|}, L_s \neq L_t, \\ 0 \qquad\qquad\qquad\quad L_s = L_t. \end{Bmatrix} \tag{9}$$

I is the intensity and σ is the intensity variance over N_s (i.e., the 8 neighbors).

3 Experiments and Results

We refer to our method as GC_{ME} (Graph Cut with Multiple Experts) and test it's performance on synthetic images and medical images from patients afflicted with CD. Our results are compared with the fused segmentations obtained using STAPLE[13], MAP-STAPLE [6], and Local MAP-STAPLE [5]. After obtaining the consensus segmentation of all images we adopt a 5 fold cross validation segmentation approach. A fully supervised RF classifier (RF-FSL) is derived from the training set (comprising the final annotations of different methods). RF-FSL calculates probability maps for each test voxel, whose negative log-likelihood is the penalty cost. The segmentation cost function is,

$$E(L) = \sum_{s \in P} -\log\left(Pr(L_s) + \epsilon\right) + \lambda \sum_{(s,t) \in N_s} e^{-\frac{(I_s - I_t)^2}{2\sigma^2}} \cdot \frac{1}{\|s-t\|}, \tag{10}$$

where $Pr(L_s)$ is the probability map of test image obtained by RF-FSL. If the training labels were obtained using GC_{ME} then the RF-FSL segmentation of the test image is compared with the ground truth segmentation from GC_{ME}. Similar tests are performed for all other label fusion methods. Each dataset was part of the test set exactly once. The method giving the most accurate consensus segmentation would result in a RF-FSL that gives the most accurate probability maps, and the corresponding segmentation would have higher agreement with the ground truth consensus segmentations. Thus the relative merit of different label fusion techniques can be judged by the accuracy of consensus segmentations obtained through them. λ was varied from $[0, 1]$ in steps of 0.001 while running our algorithm on 10 patient volumes. The maximum DM was obtained for $\lambda = 0.02$, and was the value used for all our experiments. We have 50 trees in the RF, and the maximal tree depth was fixed at 20.

3.1 Synthetic Image Dataset

Figure 1 (c) shows an example synthetic image where the 'diseased' region is within the red square. Pixel intensities are normalized to $[0, 1]$. Intensities within the square have a normal distribution with $\mu \in [0.6 - 0.8]$ and different σ. Background pixels have a lower intensity distribution ($\mu \in [0.1 - 0.3]$ and different

σ). A set of 20 adjacent boundary points are chosen and randomly displaced between $\pm 10 - 20$ pixels to obtain 3 sets of simulated segmentations (colored contours in Fig. 1 (d)). The segmentations are fused using different methods to get the final segmentation, which is compared with the reference segmentation (in Fig. 1 (c)) using Dice Metric (DM) and Hausdorff Distance (HD).

For GC_{ME} some of the expert annotations are intentionally removed to simulate real world scenarios. Variations of our method are 1) ME_{All} where all annotation information is available; 2) ME_{wSSL}, i.e., GC_{ME} without SSL for predicting missing labels. In this case the penalty costs are determined from SC_i's of available annotations. 3) ME_{wSC}, i.e., GC_{ME} without our SC score. The penalty cost is the $\chi 2$ distance between the reference distribution in the ground truth annotation of Fig. 1 (c) and the distribution from the 'expert's' annotation. Note that this condition can be tested only for synthetic images where we know the actual labels of each pixel.

Table 1 summarizes the performance of different methods. ME_{All} gives the highest DM and lowest HD values, followed by GC_{ME}, [5], [6], [13], ME_{wSSL} and ME_{wSC}. Since ME_{All} had access to all annotations, it obviously performed best. However GC_{ME}'s performance is very close and a Student $t-$test with ME_{All} gives $p < 0.042$ indicating very small difference in the two results. Importantly GC_{ME} performs much better than all other methods ($p < 0.01$).

The results show: 1) SSL effectively predicts missing annotation information since GC_{ME} has very close performance to ME_{All} and ME_{wSSL} shows a significant drop in performance from GC_{ME} ($p < 0.01$). 2) Our proposed self consistency score accurately quantifies the consistency level of each expert as is evident from the performance of GC_{ME} and ME_{wSC} ($p < 0.001$). Figure 1 (e)-(h) shows the final segmentations obtained using four different methods. The best results are obtained by GC_{ME}, followed by [5], [6] and [13].

3.2 Real Patient Dataset

3D T1-weighted spoiled gradient echo sequence (SPGE) images were acquired from 45 CD patients (excluding the 10 used to calculate λ) in supine position using a 3-T MR imaging unit (Intera, Philips Healthcare). The spatial resolution of the images was $1.02 \times 1.02 \times 2$ mm, and the acquired volume dimension was $400 \times 400 \times 100$ voxels. 2 experts annotated each slice showing CD activity. All aforementioned label fusion techniques were used to generate consensus segmentations for all patients. A 5-fold cross validation strategy was used for RF-FSL training and subsequent segmentation. The same set of features as described in Section 2.1 were used.

Table 1 summarizes the performance of different label fusion methods. The relative performance of different methods is similar to those observed for synthetic images. Note that we do not show the results for ME_{All} since all cases do not have annotations from all experts. There is also no way to know the actual labels and hence ME_{wSC} has no relevance for real medical images. Once again the importance of SSL and SC is highlighted. By incorporating these stages we achieve significant improvement in segmentation accuracy, as compared to other

Synthetic Images								Medical Images				
	ME_{All}	GC_{ME}	[5]	[6]	[13]	ME_{wSSL}	ME_{wSC}	GC_{ME}	[5]	[6]	[13]	ME_{wSSL}
DM	92.3	91.2	88.8	87.1	85.3	84.0	83.7	89.5	87.7	85.1	83.8	82.3
HD	6.1	7.4	9.0	10.1	11.9	13.5	13.9	8.2	9.8	12.0	13.9	14.7
p	.042	-	$< .01$	$< .01$	$< .01$	$< .01$	$< .001$	-	$< .01$	$< .01$	$< .01$	$< .01$

Table 1. Quantitative measures for segmentation accuracy on synthetic. DM- Dice Metric in %; HD is Hausdorff distance mm and p is the result of Student $t-$tests with respect to GC_{ME}.

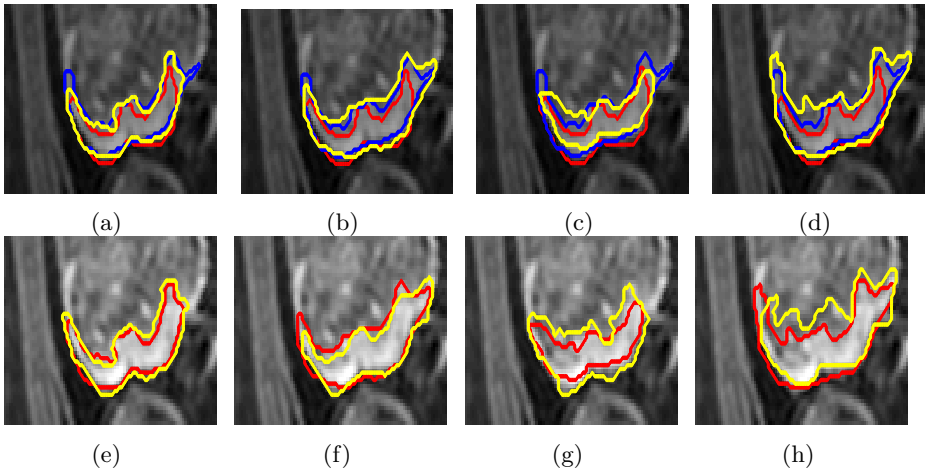

Fig. 2. First Row: The predicted ground truth for UCL Patient 23 by different methods: (a)GC_{ME}; (b) [5]; (c) [6]; and (d) [13]. Red and blue contours are expert annotations and yellow is the final annotation obtained by the respective methods **Second Row:** Segmentation results on patient 23 for: (e) GC_{ME}; (f) [5]; (g) [6]; and (h) [13]. Red contour is the ground truth segmentation while yellow contours show the final segmentation obtained by training on annotations obtained by the respective methods.

methods. Figure 2 the intermediate consensus segmentation (first row) followed by the final segmentations of the respective methods in the second row. As in the case of synthetic images, GC_{ME} gives the best results followed by [5], [6] and [13].

4 Conclusion

We have proposed a novel framework using SSL, self consistency, and GC to combine labels of multiple experts for obtaining a consensus annotation. Its performance is demonstrated by segmenting CD regions from MR images. RF based SSL classifiers predict labels of missing annotations, and avoid the iterative EM approach of other methods. Self consistency scores quantify the reliability of each expert's labels and serve as penalty costs for a second order MRF cost

function. Spatial smoothness constraints are a function of intensity difference of neighboring pixels. Experiments on synthetic and real patient datasets show the importance of our SC measure, and the effectiveness of RF-SSL in predicting missing labels. Our proposed method achieves better segmentation accuracy than competing approaches for combining multiple annotations.

References

1. Boykov, Y., Veksler, O.: Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. 23, 1222–1239 (2001)
2. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
3. Budvytis, I., Badrinarayanan, V., Cipolla, R.: Semi-supervised video segmentation using tree structured graphical models. In: IEEE CVPR. pp. 2257–2264 (2011)
4. Chatelain, P., Pauly, O., Peter, L., Ahmadi, A., Plate, A., Botzel, K., Navab, N.: Learning from multiple experts with random forests: Application to the segmentation of the midbrain in 3D ultrasound. In: In Proc: MICCAI Part II. pp. 230–237 (2013)
5. Commowick, O., Akhondi-Asl, A., Warfield, S.: Estimating a reference standard segmentation with spatially varying performance parameters: Local MAP STAPLE. IEEE Trans. Med. Imaging 31(8), 1593–1606 (2012)
6. Commowick, O., Warfield, S.: Incorporating priors on expert performance parameters for segmentation validation and label fusion: A maximum a posteriori STAPLE. In: In Proc: MICCAI Part III. pp. 25–32 (2010)
7. Criminisi, A., Shotton., J.: Decision Forests for Computer Vision and Medical Image Analysis. Springer
8. Mahapatra, D., Schüffler, P., J.Tielbeek, Makanyanga, J., Stoker, J., Taylor, S., Vos, F., Buhmann, J.: Automatic detection and segmentation of crohn's disease tissues from abdominal mri. IEEE Trans. Med. Imaging 32(12), 1232–1248 (2013)
9. Mahapatra, D., Schüffler, P., Tielbeek, J., Vos, F., Buhmann, J.: Semi-supervised and active learning for automatic segmentation of crohn's disease. In: Proc. MICCAI, Part 2. pp. 214–221 (2013)
10. Raykar, V., Yu, S., Zhao, L., Valadez, G., Florin, C., Bogoni, L., Moy, L.: Learning from crowds. Journal of Machine Learning Research 11
11. Schüffler, P., Mahapatra, D., Tielbeek, J., Vos, F., Makanyanga, J., Pends, D., Nio, C., Stoker, J., Taylor, S., Buhmann, J.: A model development pipeline for crohns disease severity assessment from magnetic resonance images. In: In Proc: MICCAI-ABD. pp. 1–10 (2013)
12. Vos, F.M., et. al.: Computational modeling for assessment of IBD: to be or not to be? In: Proc. IEEE EMBC. pp. 3974–3977 (2012)
13. Warfield, S., Zhou, K., Wells, W.: Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging 23(7), 903–921 (2004)