Obtaining Consensus Annotations For Retinal Image Segmentation Using Random Forest And Graph Cuts

Dwarikanath Mahapatra*, Joachim M. Buhmann

Department of Computer Science, ETH Zurich, Switzerland.
* *dwarikanath.mahapatra@inf.ethz.ch*

Abstract. We combine random forest (RF) classifiers and graph cuts (GC) to generate a consensus segmentation of multiple experts. Supervised RFs quantify the consistency of an annotator through a normalized consistency score, while semi supervised RFs predict missing expert annotations. The normalized score is used as the penalty cost in a second order Markov random field (MRF) cost function and the final consensus label is obtained by GC optimization. Experimental results on real patient retinal image datasets show the consensus segmentation by our method is more accurate than those obtained by competing methods.

1 Introduction

Improved algorithms have made it easier to analyse a wide variety of medical images, leading to improved computer aided diagnosis (CAD) systems. CAD systems are increasingly reliant on machine learning (ML) algorithms for image segmentation and abnormality detection. One example is glaucoma detection from retinal fundus images that requires segmentation of the optic cup (OC) and optic disc (OD) to calculate optic cup-to-disc ratio (CDR). Glaucoma is a chronic disease that affects the optic nerve resulting in its progressive damage and elongation of the optic cup. Machine learning (ML) methods for OC and OD segmentation [5] have gained importance as they provide a powerful tool for feature classification.

Success of ML based segmentation algorithms depends upon the accuracy of reference manual annotations for learning discriminative features. It is common for medical images to be manually segmented by multiple experts. A ground truth consensus segmentation is generated for validating the performance of different segmentation approaches. Obtaining consensus segmentations is challenging since manual segmentations tend to be subjective, prone to inter-observer and intra-observer variability, and of varying accuracy.

One of the first methods to combine multiple annotations, STAPLE ([12]), employed Expectation-maximization (EM) to find sensitivity and specificity values that maximize the data likelihood. Commowick et al. in [6] adapt the STAPLE algorithm to determine spatially varying performance levels using sliding windows. Chatelain et al. in [4] use Random forests (RF) to determine most

X. Chen, M. K. Garvin, J. Liu, E. Trucco, Y. Xu (Eds.): OMIA 2015, Held in Conjunction with MICCAI 2015, Munich, Germany, Iowa Research Online, pp. 41–48, 2015. Available from: `http://ir.uiowa.edu/omia/2015_Proceedings/2015/`

coherent expert decisions based on the consistency of decisions with respect to the image features but do not account for missing annotations.

The prominent challenges in obtaining a consensus annotation from multiple experts are: 1) quantifying expert's accuracy for a weighted combination of annotations; 2) predicting missing labels; and 3) ensuring spatial consistency of the final annotation. To overcome the above challenges we incorporate the the following novelties in our work: 1) an expert's reliability is quantified using supervised Random forest (RF) classifiers to generate a normalized reliability score; 2) semi supervised RF classifiers are used to predict missing labels; and 3) the normalized scores are used as penalty costs in a Markov random field (MRF) framework for spatial smoothness. The consensus annotation (ground truth) is obtained using GC optimization because: a) no iterative approach is employed as in EM based approaches of [6]; and b) globally optimum labels can be obtained thus reducing chances of getting stuck in local minima. Accuracy of consensus segmentations is validated using a ML method to segment the optic cup and disc from retinal fundus images.

2 Method

2.1 Learning Using Random Forests

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. A binary decision tree is a collection of nodes and leaves with each node containing a weak classifier that separates the data into two subsets of lower entropy. Training a node j on $S_j \subset S$ consists of finding the parameters of the weak classifier that maximize the information gain (IG_L) of splitting labeled samples S_j into S_k and S_l:

$$IG_{j,L}(S_j, S_k, S_l) = H(S_j) - \frac{|S_k}{S_j} H(S_k) - \frac{|S_l}{S_j} H(S_l) \tag{1}$$

where $H(S_i)$ is the empiric entropy of S_i. The parameters of the optimized weak classifier are stored in the node. Data splitting stops when we reach a predefined maximal depth, or when the training subset does not contain enough samples. In this case, a leaf is created that stores the empiric class posterior distribution estimated from this subset.

A collection of decorrelated decision trees increases generalization power over individual trees. Randomness is introduced by training each tree on a random subset of the whole training set (bagging), and by optimizing each node over a random subspace of the feature parameter space. At testing time, the output of the forest is defined as the average of the probabilistic predictions of the T trees.

2.2 Predicting Missing Labels

Missing labels are commonly encountered when multiple experts annotate data. We use semi-supervised learning (SSL) to predict the missing labels 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 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 for node j:

$$IG_{j,SSL} = IG_{j,UL} + \alpha IG_{j,L} \tag{2}$$

where $IG_{j,L}$ is defined in Eqn. 1. $I_{j,UL}$ depends on both labeled and unlabeled data, and is defined using differential entropies over continuous parameters as

$$I_{j,UL} = \log|\Lambda(S_j)| - \sum_{i\in\{k,l\}} \frac{|S_j^i|}{|S_j|} \log|\Lambda(S_j)| \tag{3}$$

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.

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. 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].

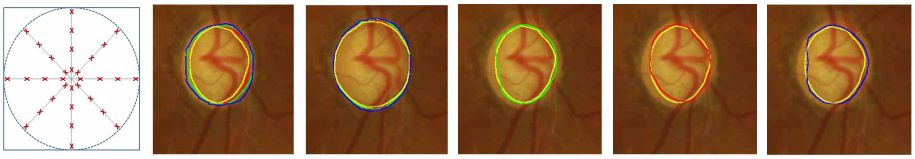
2.3 Quantifying Expert's Reliability

Expert reliability is quantified by examining the information gain at different nodes while training a random forest on samples labeled by a particular expert. This helps us evaluate the consistency of the experts with respect to the visual features. For each expert r we define an estimator $\widehat{E}_j^r$ of the expectation of the information gain on the *labeled training set* S_j sent to node j as

$$\widehat{E}_j^r = \frac{1}{\Theta_j} \sum_{\theta\in\Theta_j} IG_{j,L}^r(S_j, S_k(\theta), S_l(\theta)) \tag{4}$$

where Θ is a randomly selected subset of the feature parameters space. $\widehat{E}_j^r$ measures how well the data can be separated according to the labels of each expert. However, it suffers from two weaknesses in lower nodes of the tree: (i) it is evaluated from fewer samples, and hence becomes less reliable, and (ii) it quantifies only the experts' local consistency, without considering global consistency measures. Therefore similar to [4] we define the performance level q_j^r of each expert as a linear combination of the estimators $\widehat{E}_j^r$ from root to node j as

$$q_j^r = \frac{\sum_{d=0}^{D(j)} |S_d| \widehat{E}_{i_d(j)}^r}{\sum_{d=0}^{D(j)} |S_d|} \tag{5}$$



(a) (b) (c) (d) (e) (f)

Fig. 1. (a) template for context feature extraction; example annotations of (b) optic disc and (c) optic cup. The ground truth consensus segmentation is shown in yellow while the different expert annotations are shown in red, green and blue. Consensus segmentations for optic cup obtained using (d) GC_{ME}; (e) [6]; and (f) [1].

By weighting the estimators in proportion to the size of the training subset, we give more importance to the global estimates of the experts' consistencies, but still take into account their feature-specific performances. Once the parameters q_j^r have been computed, an expert's reliability or self consistency (SC^r) is calculated as the average performance level over all nodes j in T trees:

$$SC^r = \frac{\sum_j q_j^r}{T} \tag{6}$$

where T is the total number of trees in the forest. Higher SC^r indicates greater rater consistency.

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.

2.4 Obtaining the final labels

A second order MRF cost function is given by,

$$E(L) = \sum_{s \in P} D(L_s) + \lambda \sum_{(s,t) \in N_s} V(L_s, L_t), \tag{7}$$

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. λ 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 [2].

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 *normalized* 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.6). Since SC is higher for better agreement with labels, the corresponding penalty cost for $L_x = 1$ is

$$D(L_x = 1)^r = 1 - SC_x^r, \tag{8}$$

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

$$D(L_x = 0)^r = 1 - D(L_x = 1) = SC_x^r. \tag{9}$$

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

$$\begin{array}{l} D(L_x = 1) = \frac{1}{R} \sum_{r=1}^{R} D(L_x = 1)^r, \\ D(L_x = 0) = \frac{1}{R} \sum_{r=1}^{R} D(L_x = 0)^r. \end{array} \tag{10}$$

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

$$V(L_s, L_t) = \left\{ \begin{array}{ll} e^{-\frac{(I_s - I_t)^2}{2\sigma^2}} \cdot \frac{1}{\|s-t\|}, & L_s \neq L_t, \\ 0 & L_s = L_t. \end{array} \right\} \tag{11}$$

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 the DRISHTI-GS dataset [11]. The dataset consists of retinal fundus images from 50 patients obtained using 30 degree FOV at a resolution of 2896×1944 pixels. The optic cup and optic disc are manually segmented by 3 ophthalmologists, and the consensus ground truth is also available. We choose this dataset because the final ground truth and annotations of individual experts are publicly available and facilitates accurate validation. Quantitative evaluation is based on F-score ($F = 2\ P \times R/(P + R)$) to measure the extent of region overlap, and absolute pointwise localization error B in pixels (measured in the radial direction); P is precision and R is recall. Additionally we report the overlap measure $S = Area(M \cap A)/Area(M \cup A)$. M is the manual segmentation while A is the algorithm segmentation.

Our results are compared with the fused segmentations obtained using STAPLE[12], COLLATE [1], Majority voting (MV) and Local MAP-STAPLE [6].

After obtaining the consensus segmentations we adopt two methods to validate the accuracy of the consensus segmentation from each method. In the first method (Met 1) a separate set of fully supervised RF classifiers (RF-FSL) are trained on the consensus segmentations. The trained classifier generates probability maps for each test image voxel, whose negative log-likelihood is used as the penalty cost. The segmentation cost function is,

$$E(L) = \sum_{s \in P} -\log\left(Pr(L_s) + \epsilon\right) + \lambda \sum_{(s,t) \in N_s} e^{-\frac{(I_s - I_t)^2}{2\sigma^2}} \cdot \frac{1}{\|s - t\|}, \qquad (12)$$

where $Pr(L_s)$ is the probability map of test image obtained by RF-FSL. Since the above approach uses RFs (which is also used by GC_{ME} for predicting the fused annotations) there is the possibility of Met 1 being biased towards our method. Hence in the second validation strategy (termed Met 2) we train support vector machines (SVMs) using the same features. The trained SVM predicts the labels of each test image voxel. A convex hull is fitted to the classification map to get an initial estimate of the optic cup and disc. Subsequently active contours [10] were used to obtain the final segmentation. If the training labels were obtained using GC_{ME} then the segmentations 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. A 5 fold cross validation strategy was used for Met 1 and Met 2.

We use this validation strategy since consensus segmentations with greater accuracy are expected to give better discriminative features and the trained classifiers can identify the desired anatomy more accurately. The fusion method which most effectively combines the different annotations is expected to give highest segmentation accuracy for the test data. The relative merit of different label fusion techniques can be judged by the accuracy of consensus segmentations obtained through them. Our whole pipeline was implemented in MATLAB on a 2.66 GHz quad core CPU running Windows 7. Segmentation results on a separate dataset of 10 images gave the highest F-score for $\lambda = 0.01$ ((Eqn. 7)), which was the value fixed for our experiments. The RF has 50 trees and the maximal tree depth is fixed at 20.

3.1 Segmentation Performance

Table 1 summarizes the segmentation performance of different methods. COLLATE implementation is available from [8], while Local MAP STAPLE and STAPLE implementation is available from [9]. We closely followed the parameter setting recommendations given by the authors in the respective works. To test the SSL based prediction strategy for GC_{ME}, we create 'missing annotations' by randomly removing 1 expert's annotation for each image. We also show results for GC_{ME-All} in which none of the expert annotations were removed while predicting the final segmentation. Except for GC_{ME-All}, other methods don't have access to all annotations. Additionally, we show results for $GC_{ME-wSSL}$, i.e.,

	GC_{ME} (Met 1)	GC_{ME} (Met 2)	GC_{ME-All}	STAPLE [12]	COLLATE [1]	Local MAP STAPLE [6]	$GC_{ME-wSSL}$	Majority Voting
F	95.9	95.4	97.2	91.0	90.2	89.0	92.1	86.4
S	89.5	89.2	91.2	85.3	84.8	83.2	85.9	80.8
B	9.4	9.9	8.2	11.4	13.2	10.9	10.3	18.1
Time	7	7	7	8	6	9	7	3

Table 1. Segmentation accuracy in terms of F score, overlap and boundary distance for different methods. B is in pixels; $Time$- fusion time in minutes;F-F score; S-overlap measure; B-boundary error.

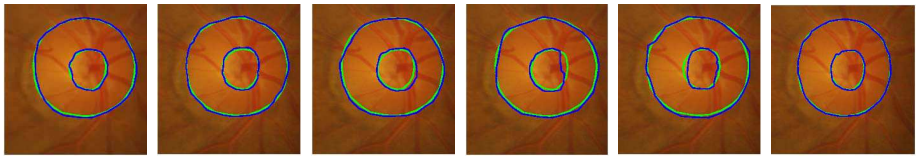
GC_{ME} without SSL for predicting the missing labels. Penalty costs are determined from SC_i's of available annotations. Missing annotations of experts is not predicted and hence not used for determining the consensus segmentation. Figure 1 (b),(c) shows the individual expert annotations and the consensus ground truth annotation while Figs 1 (d)-(f) show the predicted ground truth for 3 fusion strategies. As is evident from the images GC_{ME} shows the best agreement with the ground truth segmentations. [6] is an improved version of STAPLE.

GC_{ME} (both Met 1 and Met 1) gives the best performance among all competing methods, except GC_{ME-All}, followed by [6], [1], MV, and $GC_{ME-wSSL}$. GC_{ME}'s performance is significantly different from other methods ($p < 0.01$). Since GC_{ME-All} had access to all annotations, it obviously performed best. The results show SSL effectively predicts missing annotation information since GC_{ME} has very close performance to GC_{ME-All} ($p < 0.042$) and $GC_{ME-wSSL}$ shows a significant drop in performance from GC_{ME} ($p < 0.01$). The other important observation is that although Met 1 shows higher quantitative measures than Met 2, the difference is not significant. This is not surprising since Met 1 and the fusion strategy both use RFs. However the performance of Met 2 indicates that our fusion method is robust and performs much better than other state of the art even when not using RF classifiers for validation.

Local MAP STAPLE ([6]) shows sub-optimal performance due to predicting sensitivity and specificity parameters from annotations without considering their overall consistency. Our SC score takes into account both global and local information and is able to accurately quantify a rater's consistency. Secondly, Local MAP STAPLE may be prone to being trapped in local minima due to the iterative EM approach. On the contrary, we employ graph cuts which is almost always guaranteed to give a global minima. This makes the final output (the consensus segmentation) much more accurate and robust. COLLATE also suffers due to its reliance on iterative EM.

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 optic cup and disc from retinal images.



(a) (b) (c) (d) (e) (f)

Fig. 2. Segmentation results for different methods: (a) our proposed GC_{ME} method using Met 1; (b) GC_{ME} using Met 2; (c) [6]; (d) [1]; (e) Majority Voting; and (f) GC_{ME-All}. Green contour is manual segmentation and blue contours are algorithm segmentations from different fusion methods.

RF based SSL classifiers predict labels of missing annotations, and self consistency scores effectively quantify the reliability of each expert's labels. Graph cuts give a globally optimal solution and minimize chances of being trapped in local optima as is the case for EM based methods. Experiments show our approach outperforms other competing methods for combining multiple annotations.

References

1. Asman, A., Landman, B.: Robust statistical label fusion through consensus level, labeler accuracy, and truth estimation (COLLATE). IEEE Trans. Med. Imag. 30(10), 1779–1794 (2011)
2. Boykov, Y., Veksler, O.: Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. 23, 1222–1239 (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. Cheng, J., Liu, J., et. al.: Superpixel classification based optic disc and optic cup segmentation for glaucoma screening. IEEE Trans. Med. Imag 32(6), 1019–1032 (2013)
6. 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)
7. Criminisi, A., Shotton., J.: Decision Forests for Computer Vision and Medical Image Analysis. Springer
8. http://www.nitrc.org/projects/masi fusion/.:
9. http://www.crl.med.harvard.edu/software/.:
10. Isard, M., Blake, A.: Active Contours. Springer Verlag (1998)
11. J, S., et. al..: Drishti-gs: Retinal image dataset for optic nerve head(onh) segmentation. In: IEEE EMBC. pp. 53–56 (2014)
12. 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)