

# A CLASSIFIER ENSEMBLE BASED ON PERFORMANCE LEVEL ESTIMATION

Wei Wang<sup>a</sup>, Yaoyao Zhu<sup>a</sup>, Xiaolei Huang<sup>a</sup>, Daniel Lopresti<sup>a</sup>,  
Zhiyun Xue<sup>b</sup>, Rodney Long<sup>b</sup>, Sameer Antani<sup>b</sup> and George Thoma<sup>b</sup>

<sup>a</sup> Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA 18015

<sup>b</sup> Communications Engineering Branch, National Library of Medicine, MD 20894

## ABSTRACT

In this paper, we introduce a new classifier ensemble approach, applied to tissue segmentation in optical images of the uterine cervix. Ensemble methods combine the predictions of a set of diverse classifiers. The main contribution of our approach is an effective way of combination based on each classifier's performance level—namely, the sensitivity  $p$  and specificity  $q$ , which also produces an optimal estimate of the true segmentation. In comparison with previous work [1] that utilizes the STAPLE algorithm [2] for performance level based combination, this work achieves multiple-observer segmentation in a Bayesian decision framework using the maximum a posteriori (MAP) principle, considering each classifier as an observer. In our experiments, we applied our method and several other popular ensemble methods to the problem of detecting Acetowhite regions in cervical images. On 100 images, the overall performance of the proposed method is better than: (i) an overall classifier learned using the entire training set, (ii) average voting ensemble, (iii) ensemble based on the STAPLE algorithm; it is comparable to that of majority voting and that of the (manually picked) best-performing individual classifier in the ensemble set.

**Index Terms**— classifier ensemble, segmentation, cervigram, multiple classifier system, sensitivity, specificity

## 1. INTRODUCTION

Reliable segmentation and labeling of different regions in images are important to make images searchable by content in large medical image archives. One of the main challenges is that, due to large variations in image appearance, the color and texture feature distributions of a tissue class in one image often overlap with those of a different tissue class in other images. Therefore it is difficult to learn a single classifier that does tissue classification with low error [3].

A potential solution is to use a classifier ensemble [4, 1], which trains a set of diverse classifiers that disagree on their predictions and effectively combines the predictions in order to reduce classification error. A wide variety of classifier ensembles, including error-correcting output coding [5], bagging, and boosting [6], have been proposed with demonstrated success in reducing variance and bias. These methods differ

in the way an ensemble of classifiers is formed, by modifying the data, the learning task, or by exploiting algorithm characteristics such as randomized components and tree structures. And for combining predictions, average voting, weighted voting, and stacking [7] are commonly used.

Here we consider the problem of automatically segmenting the biomarker Acetowhite regions in an archive of 60,000 *cervigram* images of the uterine cervix, collected and digitized by NLM and NCI. Previous work on this problem has reported limited success using K-means clustering [8], Gaussian Mixture Models [9], Support Vector Machine (SVM) classifiers [3]. Although increasing the size of feature set and size of training dataset holds promise to improve performance, the intrinsic diversity in the large archive calls for ensemble methods to achieve better detection and segmentation performance (see Figure 1).

Traditional prediction combination schemes in an ensemble such as voting do not always consider the performance level of each individual classifier, therefore results may deteriorate because of poor-performing classifiers. A recent study [1] proposes a multi-classifier ensemble strategy based on the STAPLE [2] algorithm. STAPLE is a multiple-observer segmentation evaluation algorithm, which probabilistically estimates the true segmentation (ground truth map) by optimal combination of observed segmentations and a prior model of the truth. As pointed out in [10], however, in certain scenarios when the truth prior is dominant, STAPLE cannot take advantage of meaningful observer-performance priors. Shape priors are also proposed for cervigram segmentation [11]; such priors are applicable to boundaries of the cervix but not to other region boundaries such as those of Acetowhite (AW) since AW regions could be of arbitrary shape.

We propose a novel classifier ensemble based on performance level estimation and test its performance on AW segmentation in cervigrams. The overall framework is illustrated in Figure 2. A core component of the algorithm is the Multiple Observer Segmentation Evaluation system (MOSES) [10]. Unlike the Expectation Maximization based STAPLE algorithm, MOSES is a Bayesian Decision framework that computes not only a probabilistic estimate of the true segmentation but also performance measures for the individual segmentations (sensitivity  $p$  and specificity  $q$ ). The

strength of MOSES is that it effectively integrates two kinds of prior knowledge: the truth prior and the observer prior, in a balanced way so that the observer prior is properly taken into account. In the proposed ensemble algorithm, we first generate a multiple-observer ground truth map for each training and validation image using MOSES. Then a multi-label SVM classifier is learned based on each training image. The resulting SVM classifiers serve as individual classifiers (or observers) in the ensemble and their performance level estimates ( $p$  and  $q$ ) are obtained on a validation image set. Finally, the performance measures are used along with the set of individual classifiers' predictions as input to MOSES to compute the final probabilistic estimate of the true segmentation on a test image.

## 2. METHODOLOGY

In this section, we first introduce the multiple observer segmentation evaluation system (MOSES) and then describe the classifier ensemble based on performance measures.

### 2.1. Multi-Observer Segmentation Evaluation System (MOSES)

The MOSES segmentation evaluation framework takes multiple observers' segmentations and two kinds of prior knowledge as inputs: the truth prior and the observer prior, and generates as output a probabilistic estimate of the true segmentation (ground truth map). In this paper, we treat each individual classifier as an observer, and apply MOSES to forming a classifier ensemble.

Suppose there are  $N$  pixels in an image whose segmentations are being predicted by a total of  $R$  classifiers. The following notations are used in describing MOSES:

- $p = (p_1, p_2, \dots, p_R)^T$ : a column vector of  $R$  elements, with each element a sensitivity parameter characterizing one of the  $R$  segmentations;
- $q = (q_1, q_2, \dots, q_R)^T$ : a column vector of  $R$  elements, with each element a specificity parameter characterizing one of the  $R$  segmentations;
- $D$ : an  $N \times R$  matrix describing the binary decisions made for each segmentation;
- $T$ : an indicator vector of  $N$  elements, representing the hidden binary true segmentation. For each pixel  $i$ , the structure of interest is recorded as present ( $T_i = 1$ ) or absent ( $T_i = 0$ );
- $\gamma = f(T_i = 1), i = 1, \dots, N$ : the truth prior, which is the prior probability of pixel  $i$  being foreground ( $T_i = 1$ ).

The maximum a posteriori (MAP) estimator is applied to select the most probable ground truth  $T$  that maximizes the posterior distribution  $f(T|D)$ ;

$$T^* = \arg \max_T f(T|D) \quad (1)$$

For any pixel  $i$ , let

$$\begin{aligned} A_i &= f(D_{ij}|T_i = 1) \\ &= \left( \prod_{j:D_{ij}=1} p_j \prod_{j:D_{ij}=0} (1 - p_j) \right) f(T_i = 1) \end{aligned} \quad (2)$$

$$\begin{aligned} B_i &= f(D_{ij}|T_i = 0) \\ &= \left( \prod_{j:D_{ij}=0} q_j \prod_{j:D_{ij}=1} (1 - q_j) \right) f(T_i = 0) \end{aligned} \quad (3)$$

In our computation of  $A_i$  and  $B_i$  above, the observer performance priors ( $p_j, q_j, j = 1 \dots R$ ) are obtained through experiments on a validation image set, and the truth prior is initialized with uniform distribution ( $f(T_i = 1) = f(T_i = 0) = 0.5$ ). Then

$$f(T_i = 1|D) = \frac{f(D|T_i = 1)f(T_i = 1)}{\sum_T f(D|T_i)f(T_i)} = \frac{A_i}{A_i + B_i} \quad (4)$$

where  $f(T_i = 1|D)$  indicates the posterior probability of the true class label at pixel  $i$  being 1 (i.e. foreground pixel). It follows that the posterior background probability  $f(T_i = 0|D) = (1 - f(T_i = 1|D))$ . Thus the MAP estimator assigns to pixel  $i$  the class label 1 if  $f(T_i = 1|D) > 0.5$ , otherwise it assigns the label 0 (i.e. background pixel) when  $f(T_i = 1|D) < 0.5$ .

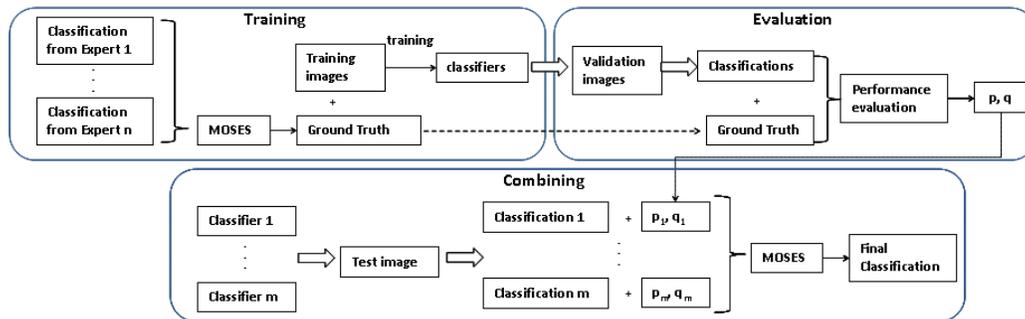
### 2.2. Training Multiple Classifiers in the Ensemble

For training purposes, we use images with region boundaries manually marked by medical experts. Since boundaries of each image are annotated by several experts, MOSES is applied to combining the multiple experts' segmentations and generating one "ground truth" segmentation. No prior knowledge is available on the level of expertise of each expert, therefore all experts are considered equally competent, with performance measures of their segmentations being ( $p = 0.9999, q = 0.9999$ ). The truth prior is initialized with a single global (homogeneous) prior as the sample mean of the relative proportion of a label in the multiple experts' segmentation data:  $\gamma = f(T_i = 1) = \frac{1}{RN} \sum_{j=1}^R \sum_{i=1}^N D_{ij}$ . Figure 1a and 1b show one example of the multi-expert segmentations and the ground truth computed by MOSES.

Given the ground truth segmentations, we learn a Support Vector Machine (SVM) [3] classifier based on pixel samples from every training image. Therefore we train  $M$  SVM classifiers given  $M$  training images. Pixels inside ground truth AW regions are given the class label 1 (foreground), and those outside are given the label 0 (background). We studied different color and texture features, labeling methods and kernels of SVM, and found that a multi-label, linear kernel SVM using L\*a\*b\* color features has sufficiently good performance. Because of large appearance variations among images, these SVM classifiers trained on single images give diverse predictions (see Figure 1c-f), which makes them suitable to be the individual classifiers in an ensemble.



**Fig. 1.** Training multiple SVM classifiers in the ensemble for AW segmentation. (a) Original image with AW boundary markings by multiple medical experts; (b) The ground truth map computed by MOSES; (c) AW segmentation by SVM classifier 1 ( $p = 0.576, q = 0.988$ ); (d) AW by classifier 2 ( $p = 0.954, q = 0.632$ ); (e) AW by classifier 3 ( $p = 0.776709, q = 0.967$ ); (f) AW by classifier 4 ( $p = 0.655, q = 0.990$ ).



**Fig. 2.** Overview diagram of the proposed multiple classifier system

### 2.3. Performance Evaluation on a Validation Dataset

The performance level of each SVM classifier is evaluated through experiments on a validation image set. Each of the  $M$  classifiers is applied to classifying all  $N$  validation images. Every pixel is classified to have either label 1 (AW) or 0 (non-AW). Sensitivity and specificity ( $p$  and  $q$ ) values are used as performance measures. Thus we have an  $M \times N$  matrix recording the  $(p, q)$  values of  $M$  classifiers on  $N$  validation images. Since a classifier ensemble using MOSES requires  $p$  and  $q$  to be scalar values, we compute the performance level of each classifier as the average  $(\bar{p}, \bar{q})$  of the classifier's  $(p, q)$  values on all  $N$  validation images. That is, for the  $i$ th classifier ( $i = 1 \dots M$ ),  $\bar{p}_i = \frac{1}{N} \sum_{j=1}^N p_{ij}$ ,  $\bar{q}_i = \frac{1}{N} \sum_{j=1}^N q_{ij}$  where  $p_{ij}$  and  $q_{ij}$  are the sensitivity and specificity, respectively, of the  $i$ th classifier on the  $j$ th validation image.

### 2.4. Classifier Ensemble

On a test image, the predictions (i.e. segmentations) of the multiple individual SVM classifiers are combined in MOSES to generate an estimate of the true segmentation (see Section 2.1). The classifier  $\bar{p}_i$  and  $\bar{q}_i$  values obtained in Section 2.3 are taken as the observer priors for MOSES. The other prior of MOSES, the truth prior  $\gamma$ , is initialized to 0.5 since on the test image we assume no knowledge about the probability of one pixel inside AW regions.

## 3. EXPERIMENTS

We implemented our algorithm in Matlab 2007b on a computer with Intel Core2 E6850 CPU. 939 cervigram images from the NCI/NLM archive with multiple-expert boundary markings are available for training and validation purposes. We used 100 images of diverse appearance for training ( $M =$

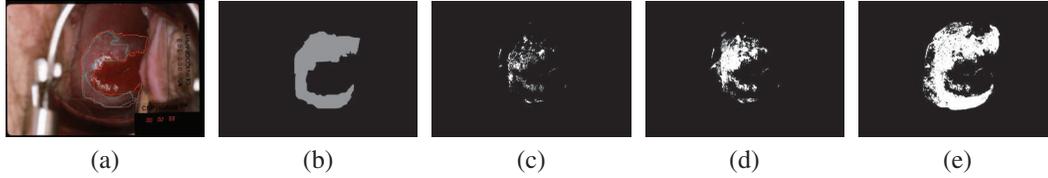
100), and another 100 images for validation and classifier performance evaluation ( $N = 100$ ). One multi-label, linear kernel SVM is learned based on each training image; the multi-label SVM classifier can simultaneously segment several important tissue regions in cervigrams, including the AW, Columnar Epithelium (CE) and Squamous Epithelium (SE). When performing AW segmentation, all other labels (CE, SE, others) are considered as background. Then a classifier ensemble is formed based on individual classifiers' performance measures, and the final segmentation on a test image is computed by the classifier ensemble through MOSES.

We first compared results of the proposed classifier ensemble with those of a single overall classifier without ensemble. To do this, we separately trained an overall SVM classifier using all 100 images. The performance measure comparison on the 100 validation images is shown in Table 1. Figure 3 visually demonstrates such comparison on a test image. One can see that the proposed ensemble classifier gives better segmentation results than the overall classifier. The ensemble also achieves a performance level similar to that of the best-performing individual classifier in the set (Table 1).

We further compared the proposed method with other classifier ensemble methods including average voting, majority voting, and STAPLE ensemble (Table 2). The proposed ensemble using MOSES compares favorably with average voting and STAPLE ensemble, and it is comparable with majority voting by individual classifiers.

## 4. CONCLUSION AND DISCUSSION

We introduce a multiple classifier ensemble approach based on performance evaluation, and apply it to segmenting tissue regions, especially the biomarker acetowhite tissue in digi-



**Fig. 3.** (a) Multi-experts' AW boundary markings, (b) Ground truth from MOSES, (c) Segmentation by the overall SVM classifier ( $p = 0.096, q = 0.945$ ), (d) Segmentation by the proposed classifier ensemble ( $p = 0.318, q = 0.986$ ), (e) Segmentation by the best individual classifier ( $p = 0.851, q = 0.726$ ).

Method	$\bar{p}$	$\sigma$ of $p$	$\bar{q}$	$\sigma$ of $q$
Overall	0.3606	0.1941	0.8873	0.1238
Proposed ensemble	0.6108	0.2076	0.7064	0.2207
Best individual	0.7129	0.1275	0.7822	0.1068

**Table 1.** Performance comparison (mean and standard deviation of  $p$  and  $q$ ) among three methods: an overall classifier learned using all training images, proposed ensemble classifier, best individual classifier in the ensemble set.

Methods	$\bar{p}$	$\sigma$ of $p$	$\bar{q}$	$\sigma$ of $q$
average voting	0.2918	0.1739	0.9017	0.1099
majority voting	0.6263	0.1935	0.6729	0.2273
STAPLE	0.3006	0.1751	0.8974	0.1121
MOSES	0.6108	0.2076	0.7064	0.2207

**Table 2.** Comparison among four different ensemble methods: average voting, majority voting, combining classifiers by STAPLE, and proposed method of combining classifiers based on performance measures and MOSES.

tized uterine cervix images. The multiple classifier system uses a multi-observer segmentation evaluation tool (MOSES) to train and combine SVM classifiers. Experimental results show that the proposed classifier ensemble performs better than a single SVM classifier learned on all training images, better than average voting as well as classifier combination using the STAPLE algorithm. It achieves comparable results with majority voting and the best individual classifier. The proposed classifier ensemble shares common weaknesses with other ensemble methods, which are increased storage and computation time. In our future work we will investigate mechanisms to make the proposed ensemble more efficient, and explore active and online learning algorithms to better solve the cervigram segmentation problem.

## 5. ACKNOWLEDGEMENTS

We would like to thank the Communications Engineering Branch, National Library of Medicine—NIH, and the Hormonal and Reproductive Epidemiology Branch, National Cancer Institute—NIH, for providing the data and support of this work.

## 6. REFERENCES

- [1] Y. Artan and X. Huang, "Combining multiple  $2\nu$ -SVM classifiers for tissue segmentation," Proc. of *ISBI* 2008, pp. 488–491.

- [2] S.K. Warfield, K.H. Zou, and W.M. Wells, III, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. on Medical Imaging*, vol. 23, no. 7, pp. 903–921, 2004.
- [3] X. Huang, W. Wang, Z. Xue, S. Antani, L. R. Long, and J. Jeronimo, "Tissue classification using cluster features for lesion detection in digital cervigrams," in *SPIE, Medical Imaging: Image Processing, 2008*.
- [4] M. De Santo, G. Percannella, C. Sansone, M. Vento, "A neural multi-expert classification system for MPEG audio segmentation," in *Advances in Pattern Recognition*, pp. 50–59, 2001.
- [5] N. Yamaguchi and N. Ishii, "Combining classifiers in error correcting output coding," *Syst. Comput. Japan*, vol. 35, no. 4, pp. 9–18, 2004.
- [6] Thomas G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Mach. Learn.*, vol. 40, no. 2, pp. 139–157, 2000.
- [7] Ludmila I. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 281–286, 2002.
- [8] B. Tulpule, D. Hernes, Y. Srinivasan, S. Yang, S. Mitra, Y. Sriraja, B. Nutter, B. Phillips, L.R. Long, and D. Ferris, "A probabilistic approach to segmentation and classification of neoplasia in uterine cervix images using color and geometric features," in *SPIE, Medical Imaging: Image Processing*, Vol. 5747, pp. 995–1003, 2005.
- [9] S. Gordon, G. Zimmerman, R. Long, S. Antani, J. Jeronimo, and H. Greenspan, "Content analysis of uterine cervix images: Initial steps towards content based indexing and retrieval of cervigrams.," in *SPIE, Medical Imaging: Image Processing*, Vol. 6144, pp. 2037–2045, 2006.
- [10] Y. Zhu, X. Huang, W. Wang, D. Lopresti, Z. Xue, R. Long, S. Antani, and G. Thoma, "Balancing the role of priors in multi-observer segmentation evaluation," *Journal of Signal Processing Systems: Special Issue on Biomedical Imaging*, May 2008.
- [11] S. Gordon, S. Lotenberg and H. Greenspan, "Shape priors for segmentation of the cervix region within uterine cervix images," in *SPIE medical imaging*, 2008.