ADJUSTABLE ADABOOST CLASSIFIER AND PYRAMID FEATURES FOR IMAGE-BASED CERVICAL CANCER DIAGNOSIS

Tao Xu[1]*, Edward Kim*[2]*, Xiaolei Huang*[1]

[1] Computer Science and Engineering Department, Lehigh University, Bethlehem, PA, USA;
[2]Department of Computing Sciences, Villanova University, Villanova, PA, USA

ABSTRACT

Cervical cancer is the third most common type of cancer in women worldwide. Most death cases of cervical cancer occur in less developed areas of the world. In this work, we develop an automated and low-cost method that is applicable in those low-resource regions. First, we propose a more distinctive multi-feature descriptor for encoding the cervical image information by enhancing an existing descriptor with the pyramid histogram of local binary pattern (PLBP) feature. Second, we apply the AdaBoost algorithm to perform feature selection, and train a binary classifier to differentiate high-risk patient visits from low-risk patient visits. Our AdaBoost classifier can be adjusted to achieve high specificity, which is necessary for use in clinical practice. Experiments on both balanced and imbalanced datasets are conducted to evaluate the effectiveness of our method. Our method is shown to achieve better performance than existing image-based CIN classification systems and also outperform human interpretations on various screening tests.

Index Terms— Cervical Cancer Screening, Computer Aided Diagnosis, Image Classification, AdaBoost, Pyramid Histograms, Local Binary Patterns

1. INTRODUCTION

Cervical cancer ranks as the second most common type of cancer in women aged 15 to 44 years worldwide [1]. Among death cases caused by cervical cancer, over 80% occurred in less developed regions. Therefore, there is a need for lower cost and more automated screening methods for early detection of cervical cancer [2], especially those applicable in low-resource regions.

Screening procedures prevent cervical cancer by detecting cervical intraepithelial neoplasia (CIN), which is the potentially precancerous changes and abnormal growths of squamous cells on the surface of the cervix. According to WHO system [1], CIN is divided into three grades: CIN1 (mild), CIN2 (moderate), and CIN3 (severe). In clinical practice, one of the most important goals of screening is to differentiate CIN1 from CIN2/3 or cancer (denoted as CIN2/3+ [3, 4]). The reason is that the lesions in CIN2/3+ require treatment, whereas mild dysplasia in CIN1 only needs conservative observation because it will typically be cleared by an immune response in a year.

The most widely used cervical cancer screening and diagnostic methods include Pap tests, HPV tests, Colposcopy, and Digital Cervicography. Pap tests are effective, but suffer from low sensitivity in detecting CIN 2/3+ [5]. Moreover, Pap tests need a laboratory and trained personnel to evaluate the samples. The sensitivity of HPV tests in detecting CIN 2/3+ lesions varies greatly [5]. Colposcopy is a diagnostic procedure that often involves setting a biopsy. Digital Cervicography is a non-invasive and low cost visual examination method which is widely accessible in resource-poor regions. It is performed by taking a photograph of the cervix (called a cervigram) after applying 5% acetic acid to the cervix epithelium. As summarized in a recent review article [6], the conventional cervical screening methods are highly dependent on the skill of the experts whose judgment may be subjective and often leads to considerable variability. Furthermore, there are far more patients than expert physicians, resulting in long queues for the screening process. Also, in developing areas of the world, patients cannot gain access to some screening tests. To overcome these problems, computational, efficient and low-cost adjunctive tools are needed for cervical cancer diagnosis.

In this paper, we present a machine learning based method to differentiate CIN1/normal from CIN2/3+ using cervigram images obtained during a visit of the patient. In our experiments, we apply our method on both balanced and imbalanced datasets to evaluate its robustness. We encode the cervigram information into a multi-feature descriptor which combines three types of complementary features: the pyramid histogram of oriented gradients (PHOG) [7], the pyramid color histogram in L*A*B space (PLAB) and the pyramid histogram of local binary pattern (PLBP). In this multi-feature descriptor, PHOG encodes edges and gradient information, PLAB captures color information and PLBP extracts texture information. Compared with only using PLAB and PHOG features [3, 4], the proposed multi-feature descriptor achieves better performance by adding PLBP, e.g., it improves the sensitivity from 74% to 86% at 70% specificity on the imbalanced dataset. But for computer-aided interpretation of cervigrams to be useful in clinical practice, it is important to develop methodologies that can achieve a specificity around 90% and a sensitivity as high as possible. Thus, we introduce a factor in the AdaBoost learning algorithm to control the trade-off between sensitivity and specificity of the classifier. The proposed adjustable AdaBoost algorithm is used to select

discriminative features and train classifiers. Compared with the support vector machine (SVM) classifier [3], this algorithm achieves higher sensitivities by using fewer attributes especially at the high specificity region, e.g., AdaBoost classifier increases the sensitivity by 8% at 90% specificity on the balanced dataset. Moreover, the experimental results illustrate that the proposed automated algorithm achieves a much higher sensitivity than the human interpretations on various screening tests, including Cervicography, Pap, and HPV tests.

2. RELATED WORKS

Several computer-assisted Pap tests have been approved by United States Food and Drug Administration (USFDA), such as ThinPrep Imaging System (TIS) [8] and FocalPoint [9]. These methods were shown to be statistically more sensitive than manual methods with equivalent specificity. Encouraged by these developments, a data-driven algorithm [3] was developed for automated cancer diagnosis via analyzing cervigrams. Compared with Pap tests [8, 9], cervigrams are images captured by the non-invasive and low cost digital cervicography. The method by Kim et al. [3] utilized a linear support vector machine (L-SVM) to learn image features and classify cervigrams into CIN1/normal or CIN2/3+. To further improve the classification performance, Song et al. [4] combined the cervigram information with other clinical test results such as Pap and HPV. However, these other clinical tests require additional resources that may not be available in resource poor areas of the world.

The choice of feature descriptors is one of the most important factors for image segmentation and classification. Several types of features [3, 4, 10–12] have been proposed to encode cervigram information. Li et al. [10] identified acetowhite regions by analyzing local color features. Zimmerman et al. [11] detected specularities in cervigrams by utilizing image intensity, saturation, and gradient information. In the work by Ji et al. [12], texture features were used to recognize important vascular patterns in cervigrams. In [3, 4], the authors combined the pyramid histogram of oriented gradients (PHOG) and the pyramid color histogram in L*A*B space (PLAB) features to perform region of interest (ROI) segmentation and CIN classification.

In addition to feature descriptors, classifiers also have great influence on the performance of a machine-learning based classification method. Neural networks, support vector machines (SVM), nearest neighbors (KNN), linear discriminant analysis (LDA), and decision trees are commonly used for studying cervical cancer [6]. Kim et al. [3] applied a linear SVM to classify cervigrams into CIN1/normal or CIN2/3+, while Song et al. [4] utilized KNN coupled with a majority voting algorithm to perform the CIN classification. Zhang et al. [13] proposed a discriminative sparse representation for tissue classification in cervigrams. In the work by Lee et al. [14], the authors developed a system which integrates multiple classifiers including neural network classifiers, statistical binary decision tree classifiers, and a hybrid classifier.

3. METHODOLOGY

Fig.1 illustrates the CIN classifier training procedure. First, we isolate the cervix region of interest (ROI) from the input image and resize it to the uniform 300*250 pixels. We use the method proposed in [3] to segment the ROI since segmentation is not the focus of this paper. Second, we transform the ROI image patch into different types of feature maps, including the local binary pattern (LBP) map, L*A*B color channels, and the image gradient maps. Third, a spatial pyramid of sub-regions is constructed for each feature map. Based on these constructed pyramids, PLBP, PLAB and PHOG features are extracted and concatenated to be a multi-feature descriptor. Finally, the adjustable AdaBoost algorithm is applied to select discriminative attributes and train a CIN classifier on the multi-feature descriptor.

3.1. Image to Feature Map Transformation

Color and Image Gradient. Color plays an important role in cervical lesion classification, because one of the most important visual features on the cervix that have relevant diagnostic properties is the presence of Acetowhitened regions. Thus, the color feature is widely used in cervigram analysis [3, 4, 13]. We calculate the L*A*B color channels as our color feature maps. Then, to capture edge and shape information on a cervix, we calculate the gradient map, which is shown to be complementary to the color feature [3, 4].

Texture. In addition to the color and gradient features, we introduce a new local binary pattern (LBP) feature that extracts local texture characteristics for cervical lesion classification. Ojala et al. [15] first introduced LBP and showed its powerful ability for texture classification. In a local neighborhood of an input image, given a pixel (x_c, y_c) which is surrounded by 8 neighbors, we can calculate its LBP value by Eq. (1),

$$LBP(x_c, y_c) = \sum_{p=0}^{7} s(i_p - i_c)2^p \quad (1)$$

Where i_c indicates the grayscale value of the center pixel (x_c, y_c); i_p corresponds to the grayscale value of the pth neighbor. $s(x)$ is a sign function where $s(x) = 1, if\ x \geq 0; else,\ s(x) = 0$.

Later, several extensions of the original LBP operator were presented [16]. First, the LBP was extended to a circular neighborhood of different radii, notated as $LBP_{P,R}$ which refers to P equally spaced pixels on a circle of radius R. Furthermore, the rotation invariant local binary pattern is

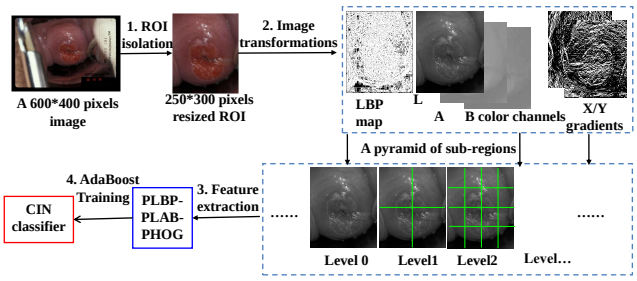

Fig. 1. The CIN classifier training framework

defined in Eq. (2),

$$LBP^{ri}_{P,R} = \min_i ROR(LBP_{P,R}, i), i = 0, ..., P-1 \quad (2)$$

Where $ROR(LBP_{P,R}, i)$ performs a circular bitwise right shift on the P-bit $LBP_{P,R}$, for i number of times.

By replacing the intensity value i_c of each pixel (x_c, y_c) in the input image with the $LBP^{ri}_{P,R}$ value, we can obtain the LBP map. The pixels on the boundary of the input image do not correspond to any LBP values, thus we need to set the values of those pixels to be zeros or to be the values of their closest neighbors that have LBP values.

In this paper, we use $LBP^{ri}_{8,1}$. There is no need to use LBP with other radii because our pyramid histogram LBP feature (PLBP) can encode a multi-scale local binary pattern.

3.2. Pyramid Feature Extraction

As Fig.1 shows, we need to construct a spatial pyramid for each feature map. A pyramid is constructed by splitting the image into rectangular sub-regions, increasing the number of regions at each level, i.e., level 0 has 1 sub-region; level 1 has 4 sub-regions; level 2 has 16 sub-regions, and so forth. Histogram features are extracted within these pyramid sub-regions. The extracted pyramid histogram encodes the statistical distribution of feature values at different positions and scales in cervigrams.

For the PLBP feature, the total number of bins is 10 for the histogram of a subregion. A 4-level of pyramid is constructed resulting in a PLBP histogram feature that has 850 dimensions. For the PLAB feature, we extract 3 pyramid levels with a 16-bin histogram for each channel in L*A*B color space in each subregion. Thus, the PLAB color feature has 1,008 dimensions. In the gradient map, we calculate pyramid histogram of oriented gradients (PHOG). An 8-bin orientation histogram over 4 levels is used. Hence, the total vector size of our PHOG feature is 680. Finally, we construct a multi-feature descriptor by concatenating the three different types of features, PLBP-PLAB-PHOG. Thus, this multi-feature descriptor has a vector size of 2,538.

3.3. Adjustable AdaBoost Classifier

Boosted decision trees are frequently used to select a discriminative feature subset from a feature pool and train a classifier. In this paper, we utilize a quickly boosting learning algorithm [17] to train an AdaBoost classifier on our PLBP-PLAB-PHOG feature descriptor.

A boosted classifier has the form $H(x) = \sum_t \alpha_t h_t(x)$, which can be trained by greedily minimizing the loss function and selecting the weight scalar α_t and optimal weak classifier h_t at each training iteration t. We use shallow decision trees (i.e. stumps) as the weak learners. The decision tree $h_t(x)$ is composed of a stump at every non-leaf node. A stump is trained on a single feature and determines an optimal threshold on this feature's value that minimizes the loss function using this feature. In the t-th iteration, among all stumps, the stump that gives minimum error is continuously selected to construct the weak classifier $h_t(x)$. In the final strong classifier $H(x)$, the weight of weak classifier $h_t(x)$ is α_t, which is inversely proportional to the classification error of $h_t(x)$.

In this paper, since we classify each patient visit and there are often multiple images taken during a visit, the final classification label of a visit is determined by considering the classification results on all images of this visit. Let $x_1, ..., x_m$ be the multi-feature descriptors for m images of a patient visit. The final label of this patient visit is determined by Eq. (3),

$$L(x) = sign(\sum_{i=1}^{m}\sum_{t=1}^{T} \alpha_t h_t(x_m) - \delta) \quad (3)$$

Here, we introduce a factor δ to control the trade-off between sensitivity and specificity of the final classifier. $\delta \in [0, A]$, with default value $\frac{A}{2}$, where $A = m\sum \alpha_t$. By increasing δ, the classifier achieves higher specificity with lower sensitivity. We add this factor here because in clinical practice, it is often desired to ensure the specificity of a screening test above 90% while achieving a sensitivity as high as possible.

4. EXPERIMENTS

We carry out our experiments using data from a database collected by the NCI (National Cancer Institute) in the Guanacaste project [18]. Cervigrams and other clinical information of 10,000 anonymized women are available. Since our goal is to perform visit level CIN classification based on image information, the ground truth used for validation is the worst histology of each patient visit, obtained from microscopic evaluation of tissue samples taken during biopsy. Since the Guanacaste project was a population-based study of cervical neoplasia, a large number of women who were screened were healthy. From the database [18], we select 1,112 patient visits which have worst histology ground truth information. Among these, 767 visits are in the CIN1/normal category and 345 visits are in the CIN2/3+ category. For our experiments, we construct two datasets. D^{imb} contains all 1,112 patient visits; it is an imbalanced dataset with an imbalance ratio of 2.22:1 (negative:positive). Meanwhile, in order to compare our method with existing methods which reported results only on balanced datasets [3, 4], we construct a balanced dataset D^b by using all 345 CIN2/3+ visits and 345 randomly selected CIN1/normal visits.

We perform a ten-round ten-fold cross validation on both D^b and D^{imb} datasets to evaluate the sensitivity and specificity of our method. We randomly divide the visits in each dataset into ten folds. In the ten rounds, we rotationally use one fold for testing, one fold for validation and the remaining eight folds for training. We report the average result of the ten rounds. In our AdaBoost learning algorithm, the only parameter (the number of weak classifiers) can be automatically learned by optimizing the GMean on the validation set, where $GMean = \sqrt{sensitivity * specificity}$. In the testing process, by changing δ, we can draw ROC curves for AdaBoost classifiers trained on different features. As the baseline for comparison, we use the source code of the method proposed in paper [3] received from its authors. For fair comparison,

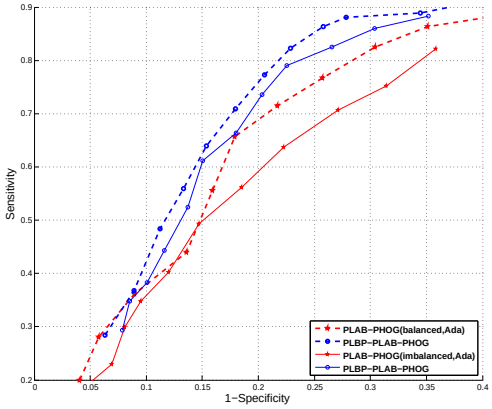

Fig. 2. Comparison of AdaBoost Classifiers trained on PLBP-PLAB-PHOG and PLAB-PHOG feature descriptors. (Bold dash lines show results on balanced dataset; Thin solid lines show results on imbalanced dataset; Red lines are results based on PLAB-HOG (baseline); Blue lines are results based on PLBP-PLAB-PHOG.)

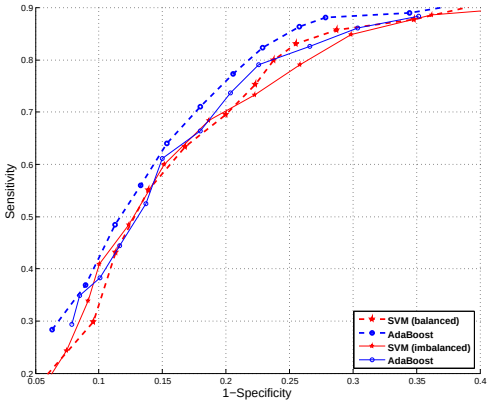

Fig. 3. Comparison between AdaBoost and SVM classifiers trained on PLBP-PLAB-PHOG. (Bold dash lines show results on balanced dataset; Thin solid lines show results on imbalanced dataset; Red lines are results of SVMs; Blue lines are results of AdaBoosts.)

we train and test their method on our datasets under the same condition as our method. We also compare our result with human interpretations on several other screening tests, obtained for the same visits that are used to construct our datasets.

4.1. PLBP-PLAB-PHOG Feature vs. PLAB-PHOG

In this experiment, we evaluate the performance of our PLBP-PLAB-PHOG feature descriptor by comparing it with the baseline feature PLAB-PHOG [3, 4] on both balanced and imbalanced datasets. In Fig. 2, we report results in ROC curves produced by the adjustable Adaboost classifier trained on different features. The δ factor in Eq. (3) is varied from -0.12 to 0.08 to generate data points on the ROC curves. As Fig. 2 shows, the PLBP-PLAB-PHOG feature (blue lines) outperforms PLAB-PHOG (red lines) on both datasets, which demonstrates that adding PLBP makes a better feature descriptor for cervigram images. On the balanced dataset, our PLBP-PLAB-PHOG increases the sensitivity from about 38% to 42% at 90% specificity. Furthermore, on the balanced dataset the best accuracy of PLBP-PLAB-PHOG feature is 80.30% achieved at 86.39% sensitivity and 74.21% specificity, while the best accuracy of PLAB-PHOG is 76.10%.

4.2. Adjustable AdaBoost Classifier vs. SVM

To illustrate the performance improvement by our adjustable AdaBoost classifier, we compare it with an SVM classifier. Both types of classifiers are trained on the same dataset based on PLBP-PLAB-PHOG feature. As Fig. 3 illustrates, the overall result of AdaBoost and SVM classifiers are very close on the imbalanced dataset. However, at high specificity AdaBoost achieves higher sensitivity on both datasets. For instance, AdaBoost improves the sensitivity from about 34% to 42% at 90% specificity in comparison to SVM on the balanced dataset. Also, the overall improvement on the balanced dataset is obvious. Hence our AdaBoost classifier performs better than SVM, especially in a clinical practice scenario where a specificity of 90% or above is generally required.

Table 1. Comparing the sensitivity (Sensi) and Specificity (Speci) of our proposed AdaBoost classifier trained on PLBP-PLAB-PHOG feature with that of human interpretation (HI) on Pap and HPV tests.

Method	Sensi(%)	Speci(%)	Sensi(%)	Speci(%)
HI-Cervigram	21.74	93.04	21.74	94.52
HI-Alfaro ThinPrep	20.69	81.82	20.69	85.27
HI-Cytyc ThinPrep	49.55	88.46	49.55	89.77
HI-Costa Rica Pap	39.42	88.12	39.42	89.31
HI-Hopkins Pap	36.00	97.11	36.00	97.13
HI-HPV16	33.82	94.19	33.82	92.94
HI-HPV18	08.16	97.97	08.16	98.17
Proposed	42.00	90.00	44.40	88.37

4.3. Proposed Classifier vs. Human Interpretation

As illustrated in Table 1, on the balanced dataset the AdaBoost classifier trained on PLBP-PLAB-PHOG feature outperforms human interpretations on ThinPrep and Costa Rica Pap tests at the specificity level around 90%. When compared to all other tests except Cytyc ThinPrep, the AdaBoost classifier achieves lower specificity, but its sensitivity is much higher. Furthermore, our AdaBoost classifier is able to achieve an even better overall accuracy regardless the consideration of high specificity. For example, we can achieve the best accuracy of 80.30% with 86.39% sensitivity and 74.21% specificity on the balanced dataset. For further comparison, in [13] the best reported result is 71.15% sensitivity with 81.67% specificity. Therefore, we conclude that our AdaBoost classifier trained on PLBP-PLAB-PHOG feature can perform comparably or better than human interpretation and some state-of-the-art automated methods [3, 13].

5. CONCLUSION

We present an adjustable AdaBooast classifier for CIN classification of patient visits using image information only. Moreover, a multi-feature descriptor, PLBP-PLAB-PHOG, is designed to encode the color, edge, shape and texture information in cervigrams. Experiments on both balanced and imbalanced datasets are conducted to evaluate the effectiveness of our method. Compared with the PLAB-PHOG feature used in [3, 4], our PLBP-PLAB-PHOG feature achieves much bet-

ter result. Also, our AdaBoost classifier trained on PLBP-PLAB-PHOG feature outperforms human interpretations on various screening tests and some state-of-the-art computer assisted cervigram analysis approaches [3, 13].

References

[1] WHO/ICO Information Centre on HPV and Cervical Cancer, "Human papillomavirus and related cancers in world," in *Summary report*, Aug. 2014.

[2] P. Malm, "Image analysis in support of computer-assisted cervical cancer screening," in *Uppsala Dissertation*, 2013.

[3] E. Kim and X. Huang, "A data driven approach to cervigram image analysis and classification," *Color Medical Image analysis*, vol. 6, pp. 1–13, 2013.

[4] D. Song, E. Kim, X. Huang, and et al, "Multi-modal entity coreference for cervical dysplasia diagnosis," in *Medical Imaging*. IEEE, 2014.

[5] R. Sankaranarayanan, L. Gaffikin, M. Jacob, and et al, "A critical assessment of screening methods for cervical neoplasia," *IJGO*, vol. 89, pp. 4–12, 2005.

[6] Y. Jusman, S.C. Ng, and N.A.A. Osman, "Intelligent screening systems for cervical cancer," *The Scientific World Journal*, 2014.

[7] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *CIVR*, 2007, pp. 401–408.

[8] C. V. Biscotti, A. E. Dawson, and et al., "Assisted primary screening using the automated thinprep imaging system," in *AJCP*, 2005, vol. 123(2), pp. 281–287.

[9] D. C. Wilbur, W. S. Black-Schaffer, and et al., "The becton dickinson focalpoint gs imaging system: Clinical trials demonstrate significantly improved sensitivity for the detection of important cervical lesions," in *AJCP*, 2009, vol. 132(5), pp. 767–775.

[10] W. Li, J. Gu, D. Ferris, and A. Poirson, "Automated image analysis of uterine cervical images," in *SPIE Medical Imaging*, 2007.

[11] G. Zimmerman-Moreno and H. Greenspan, "Automatic detection of specular reflections in uterine cervix images," in *SPIE Medical Imaging*, 2006.

[12] Q. Ji, J. Engel, and E. Craine, "Classifying cervix tissue patterns with texture analysis," in *Pattern Recognition*, 2000, vol. 33(9), pp. 1561–1574.

[13] S. Zhang, J. Huang, and et al, "Discriminative sparse representations for cervigram image segmentation," in *ISBI*. IEEE, 2010, pp. 133–136.

[14] JS-J. Lee, J. Hwang, and et al, "Integration of neural networks and decision tree classifiers for automated cytology screening," in *IJCNN*, 1991, vol. 1, pp. 257–262.

[15] T. Ojala, M. Pietikinen, and D. Harwood, "A comparative study of texture measures with classification based on feature distributions," in *Pattern Recognition*, 1996, vol. 29, pp. 51–59.

[16] T. Ojala, M. Pietikinen, and T. Menp, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," in *PAMI*. IEEE, 2002, vol. 24, pp. 971–987.

[17] R. Appel, T. Fuchs, P. Dollr, and P. Perona, "Quickly boosting decision trees pruning underachieving features early," in *ICML*, 2013.

[18] R. Herrero et al., "Design and methods of a population-based natural history study of cervical neoplasia in a rural province of costa rica," in *The Guanacaste Project*, 1997, vol. 1, pp. 362–375.