AUTOMATIC QUANTIFICATION AND CLASSIFICATION OF CERVICAL CANCER VIA ADAPTIVE NUCLEUS SHAPE MODELING

Hady Ahmady Phoulady[1], *Mu Zhou*[2], *Dmitry B. Goldgof*[1], *Lawrence O. Hall*[1], *Peter R. Mouton*[1]

[1]University of South Florida, Tampa, FL
[2]Stanford University, Stanford, CA

ABSTRACT

Decisions about cervical cancer diagnosis and classification currently require microscopic examination of cervical tissue by an expert pathologist. In the present study, which focused on full automation of this approach, we solely use nucleus-level features to classify tissues as normal or cancer. We propose Adaptive Nucleus Shape Modeling (ANSM) algorithm for nucleus-level analysis which consists of two steps to capture the nucleus-level information: adaptive multilevel thresholding segmentation; and shape approximation by ellipse fitting. After applying the proposed algorithm, the features are extracted for tissue classification. Experiments show that ANSM can achieve an accuracy of 93.33% with a false negative rate of zero in classifying cancer and healthy cervical tissues using nucleus texture features. This provides evidence that nucleus-level analysis is valuable in cervical histology image analysis.

1. INTRODUCTION

Quantitative analysis in cervical cancer using pathological images has been a central focus for early cancer diagnosis and prevention, where pathological images provide a powerful tool to comprehensively capture tissue-level characteristics. The manual diagnosis of diseases from histology or cytology images is costly and time consuming. Potential misdiagnoses may arise from fatigue or knowledge of the expert [1], leading to over-treatment and missed disease.

Diagnostic classification of malignant and benign tissue from pathology images provides information to the pathologist for precise cancer staging assessment, accelerating patient-specific medical care. However, phenotypic niche areas including nucleus, fuzzy cell borders, and cytoplasm pose a significant barrier for evaluating slice-based malignancy.

Prior studies mostly focused the whole-slide level evaluation [2, 3] that do not consider nucleus-level features. The nucleus-level analysis from pathology images currently uses the human visual system without quantitative measurement [4]. Furthermore, current human assessment is largely restrained to semantic descriptions of size, thickness, or pleomorphism that could not substantially quantify various nucleus characteristics. In view of these challenges, we ask the following specific questions: 1) How can we design a computational framework that is capable of capturing quantitative information from the *nucleus* – instead of *whole-slide pathology images* – that is advantageous in providing useful clinical interpretation for patients? 2) If the goal is to classify tissues, do we need the most accurate segmentation or a simpler segmentation method may be as effective for classification?

In this paper, we study the problem of cervical cancer diagnostic classification based on pathology images. We suggest capturing the nucleus-level dynamics by proposing Adaptive Nucleus Shape Modeling (ANSM) algorithm to capture nucleus-level information from pathology images, including an adaptive nucleus segmentation and nucleus shape appropriation. The nucleus-level texture feature representation is then applied to classify the whole tissue malignancy. By comparing the classification accuracy using the features extracted by the proposed segmentation method to the results obtained by one of the state-of-the-art segmentation methods [5], we show that features extracted from a subset of segmented nuclei using a simpler segmentation method can be more effective for classification.

Contributions. We introduce an algorithmic framework of nucleus-level analysis to classify cervical tissues beyond conventional whole-slide analysis. Our methodological contribution is three-fold:

- We propose ANSM to capture nucleus shape in pathology images, including adaptive nucleus segmentation using a multilevel thresholding scheme and nucleus shape approximation by an ellipse shape fitting.
- We overcome the limitation of missed nuclei labeling information by proposing two intermediate steps to potentially remove poorly segmented nuclei and reduce the number of mislabeled training instances to boost classifier performance.
- We show that nucleus-level texture features obtained from segmented nuclei are effective in classifying the whole cervical tissue malignancy, providing evidence that nucleus-level analysis is valuable in understanding cervical tissue characteristics.

Related Work. Computer-aided diagnosis systems for histology analysis were proposed to classify tissue images [2] or subregions within the whole slide [6]. This is normally performed in two main steps: feature extraction and classification. Segmentation can also be used to create masks formed from patches aligned to nuclear centers [7]. Without using a proper segmentation, composite hashing and bag of features can be used to extract features [2, 8]. Popular methods to classify natural scenes [9, 10] may also be used for histology classification [7]. However, most of these systems are designed for tissue classification including lung [2, 3], breast [11, 12], prostate [6] and kidney [13] tissues and limited effort has been put on cervical tissues quantification and classification.

In [2], lung microscopic tissue images are classified as adenocarcinoma or squamous carcinoma using a Composite Anchor Graph Hashing algorithm with average accuracy 87.5%. In [11], breast histology images are separated into three regions based on blob density and classification. In [6], subregions in prostate tissue images are classified into stroma, normal or prostatic carcinoma using morphological characteristics and texture features with a classification accuracy of 79.3%. Biologically interpretable shape-based features and a series of SVM classifiers are used for classification of histological renal tumor images into three types of renal cell carcinoma and one benign tumor with an average accuracy of 77% [13]. Finally, in [14], mean nuclear volume of segmented nuclei within cervical histology images were used to classify cervical tissues with an average accuracy of 84.3% and a rejection rate of 13%.

2. ADAPTIVE NUCLEUS SHAPE MODELING

Given a cervical tissue histology image, our goal is to approximate the nucleus location and shape boundary for effective nucleus-level analysis. In this work, we made use of multilevel thresholding to accomplish nucleus segmentation, and used an ellipse shape fitting model to maximally capture the nucleus information (Fig. 1). Details are given in the following.

Adaptive Image Segmentation. We developed an adaptive multilevel thresholding nucleus segmentation method to identify the nuclei area from a raw pathology image. Due to the various nuclei morphology, it is challenging to use a single thresholding scheme for precisely delineating a nucleus shape from pathology images. To maximally preserve the nucleus geometrical shape information, while minimizing potential shape outliers, we propose a framework to capture nuclei using a segmentation algorithm described in Algorithm 1.

The segmentation algorithm can be understood as an iterative process: at each round, the darkest region R was obtained from a multilevel thresholding [15]. We incorporated morphological operations to retain the primary blobs as S_i and by comparing the total nucleus area of S_i, A_i, we measured the *goodness* of S_i. Algorithm takes the smallest and largest permitted nucleus size as input, which are denoted by m and M, respectively.

Algorithm 1 Adaptive Image Segmentation (m, M)

Set $n = 0$, $A_0 = \epsilon$
while $A_n > 0$ **do**
  Set $n = n + 1$
  Perform $n-$level thresholding and let R denote the darkest region
  Perform morphological operations filling and opening on R
  Remove blobs larger / smaller than M / m in R and denote the new region by S_n
  for each segmented blob b **do**
    Segment b using two-level thresholding and remove it if it contains more than one region larger than m
  end for
  Set A_n as the area of S_n
end while
Return S_k as the final segmentation where $k = \arg\max_i A_i$

Nucleus Shape Approximation by Ellipse Fitting. The blobs in the final segmentation were mostly a rough approximation of their exact nuclei area. Because a nucleus normally has an elliptic shape, we choose to approximate each nucleus area with the ellipse with the same normalized second central moment as the segmented nucleus area. For each nucleus region, the coordinates of the centroid of the region, with pixel coordinates (x_i, y_i) for $i \in \{1, 2, \cdots, N\}$, are

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N} \text{ and } \overline{y} = \frac{\sum_{i=1}^{N} y_i}{N}. \quad (1)$$

The central moments of order $p + q$ of a continuous bivariate probability distribution $f(x, y)$ about the mean $\mu = (\mu_X, \mu_Y)$ was defined as

$$\mu_{p,q} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \mu_X)^p (y - \mu_Y)^q f(x, y)\, dx\, dy. \quad (2)$$

Therefore, for the discrete case of binary region pixels, if we consider each pixel as a square with unit length, $f(x_i, y_i) = 1$ and the central moments are

$$\mu_{p,q} = \sum_{i=1}^{N} \left(\int_{y_i - \frac{1}{2}}^{y_i + \frac{1}{2}} \int_{x_i - \frac{1}{2}}^{x_i + \frac{1}{2}} (x - \overline{x})^p (y - \overline{y})^q\, dx\, dy \right). \quad (3)$$

Specifically, the second central moments, $\mu_{2,0}, \mu_{1,1}$ and $\mu_{0,2}$

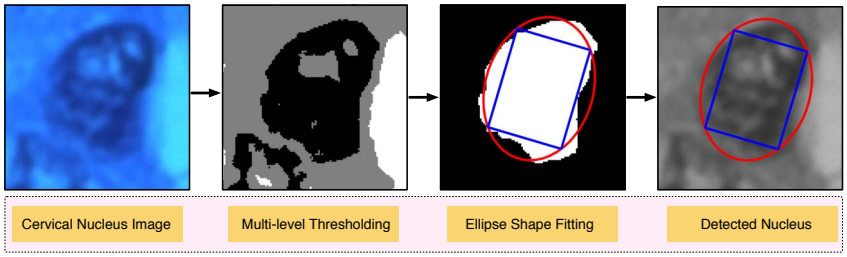

Fig. 1. An overview of the proposed Adaptive Nucleus Shape Modeling (ANSM).

are respectively computed as

$$\sum_{i=1}^{N}\left[(x_i-\overline{x})^2\right]+\frac{N}{12},\quad \sum_{i=1}^{N}\left[(x_i-\overline{x})+(y_i-\overline{y})\right],$$
$$\sum_{i=1}^{N}\left[(y_i-\overline{y})^2\right]+\frac{N}{12}, \quad (4)$$

and finally the normalized second central moments, $\mu'_{p,q}$, are defined as the second central moments divided by the number of pixels, N. Subsequently, the major axis, minor axis and orientation of the ellipse, and then coordinates of the rectangle inscribed in the ellipse were computed from the normalized central moments. This approximation using the best fit ellipse was designed to improve the area estimation for the cells with nonuniform intensity. The feature extraction was then proceeded from the maximum rectangle inscribed in the defined ellipse.

3. EXPERIMENTS AND DISCUSSION

Dataset. The dataset included 20 normal and 19 cancer sample tissues (cases). The tissues were stained with Hematoxylin and Eosin. Sample regions with normal or cancer cells were indicated on the glass slide by a pathologist and from the marked regions for each case, 10 images with size 1200x800 were acquired using a 40x objective.

Feature Extraction and Classification. The maximum rectangle inscribed in the obtained ellipse was rotated and then resized to a square with a fixed size (32x32 in this study). The shape normalization allowed for extracting the same dimension of Histogram of Oriented Gradients (HOG) or Local Binary Patterns (LBP) features from nuclei with different sizes. Instead of original HOG features [16], we used the UoCTTI variant [17] that compresses the 36 features into 31 features. The cell-size was set to 16, which decomposes the nucleus into 4 subregions, and a total number of 124 features were extracted. All the segmented nuclei in each training tissue were labeled as the label of their corresponding tissue. Segmented nuclei in testing tissues were classified by the Support Vector Machine (SVM) classifier with the Radial Basis Function (RBF) kernel and k-Nearest Neighbor (kNN) classifier in separate experiments and tissue classification was done based on the majority class of its labeled nuclei.

Second Version. Due to our focus on nuclei-level analysis, nuclei segmentation can directly affect the classification performance. Our observations indicated that, when segmentation inaccuracies occur, images of normal tissues were typically under-segmented and the images from cancer tissues were typically over-segmented. The reason was apparent on inspection: normal nuclei appear solid with uniform staining intensities. If they were isolated from other nuclei, their areas were segmented accurately. If they overlapped with other nuclei, however, they were segmented together with those nuclei, leading to under-segmentation of the nucleus areas. In contrast, cancer nuclei have non-uniform staining intensities and exhibit more heterogeneous clumping within the nuclei, that in many cases cause the segmented nucleus to be separated and treated as two or more nuclei (Fig. 2). To address this issue, in a second version of the proposed model, denoted by ANSM_2 (we denote the first version by ANSM_1), we only extracted and used features from half of the segmented nuclei: Segmented nuclei with corresponding inscribed rectangle size in the second and third quartile are kept and others are rejected.

Fig. 2 shows examples of images from normal and cancer tissues, their segmentations, nucleus approximation and rejected segmented nuclei.

Parameter Selection and Results. The only parameters used in the proposed segmentation method, m and M, were set to 250 and 5000, according to image size. To find the parameters of the SVM kernel, C and γ, a grid search was performed for selecting the best set of parameters based on the accuracy obtained from a 10-fold cross validation. Similarly, we set parameter k for kNN. The accuracy of these sets of parameters were estimated again by the average of ten, 10-fold cross validations with several sets of parameters reporting closely similar results. The reported results in Fig. 3 were obtained by setting $C = 128$ and $\gamma = 0.25$ for SVM and $k = 5$ for kNN. The highest accuracy was obtained by ANSM_2 using HOG features and SVM. For this settings, an accuracy of 93.33% was obtained, all misclassified tissues were normal tissues and therefore, false negative rate was zero. False posi-

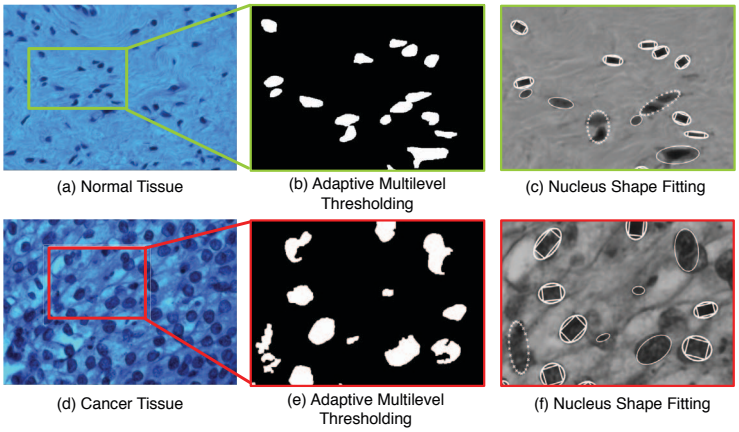

Fig. 2. Segmentation of a normal (a) and a cancer tissue (d); darkest class after multilevel thresholding and performing morphological operations (b, e); cells shown with rectangular areas are kept and the others are rejected (c, f).

tive rate was 13%, average precision through all cross validations was 87.96% and average recall was 100%.

For comparison purposes, we experimented with one of the state-of-the-art segmentation methods [5] (denoted by LTDC). LTDC needs training data to build its model. Therefore, one image from each of the cases was manually annotated and the model was trained using the 39 annotated images. The rest of the images in each case were used for testing. Also common parameters, such as m and M, which are also used by LTDC are set as in our proposed method. The results of this method and both versions of ANSM are presented in Fig. 3.

Interestingly, ANSM_1 behaves very different than ANSM_2 while ANSM_2 behaves very similar to LTDC. For example, kNN and LBP features performed better than SVM and HOG features in ANSM_1 although it was exactly the other way for both ANSM_2 and LTDC. In fact, when removing half of the training and test instances, which contained most of the misclassified labels and instances, ANSM_2 could slightly outperform LTDC and behaved very similar to it. It suggests that, for the task of classification, choosing a subset of well-segmented regions by a simpler segmentation method can be as useful as using a more advanced segmentation method.

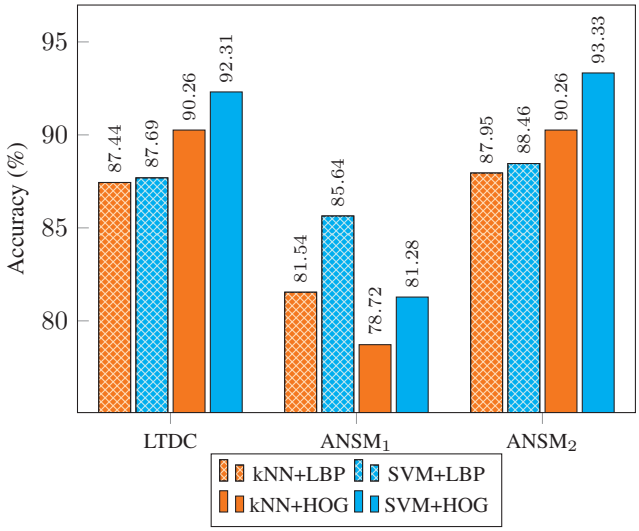

Fig. 3. Tissue classification accuracy.

4. CONCLUSIONS AND FUTURE WORK

We propose a novel algorithmic framework to tackle the challenging problem of nucleus-level pathological image analysis for cervical tissue classification. We demonstrated that the texture features extracted from segmented nuclei are able to capture class-specific tissue characteristics, which opens the space for exploring nucleus-level analysis in pathological image evaluation. Experimental results showed that our method achieved classification accuracy of 93.33% with false negative rate of zero. By comparing classification accuracy obtained using our segmentation method we showed that by using a proper shape modeling a simpler segmentation method can be as effective as a more advanced segmentation method, for the task of classification. Also, the proposed segmentation method is much faster than LTDC and does not need any training data and that makes it more applicable in real life situations. Our ongoing studies are testing this and other automatic classification algorithms in cervical tissue and cytology (Pap smears) samples immunostained with specific markers for cell proliferation and cancer-associated proteins.

5. ACKNOWLEDGEMENT

We would like to thank Erin M. Siegel and Ardeshir Hakam from H. Lee Moffitt Cancer Center and Research Institute for providing the dataset in this study.

6. REFERENCES

[1] L. J. van Bogaert, "Influence of knowledge of human immunodeficiency virus serostatus on accuracy of cervical cytologic diagnosis," *Cancer Cytopathology*, vol. 122, no. 12, pp. 909913, 2014.

[2] X. Zhang, L. Yang, W. Liu, H. Su, and S. Zhang, "Mining histopathological images via composite hashing and online learning," in *MICCAI 2014*, Polina Golland, Nobuhiko Hata, Christian Barillot, Joachim Hornegger, and Robert Howe, Eds., vol. 8674 of *LNCS*, pp. 479–486. Springer International Publishing, 2014.

[3] C. W. Wang and C. P. Yu, "Automated morphological classification of lung cancer subtypes using h&e tissue images," *Machine Vision and Applications*, vol. 24, no. 7, pp. 1383–1391, 2013.

[4] M. T. McCann, J. A. Ozolek, C. A. Castro, B. Parvin, and J. Kovačević, "Automated histology analysis: Opportunities for signal processing," *Signal Processing Magazine, IEEE*, vol. 32, no. 1, pp. 78–87, Jan 2015.

[5] Carlos Arteta, Victor Lempitsky, J Alison Noble, and Andrew Zisserman, "Learning to detect cells using non-overlapping extremal regions," in *Medical image computing and computer-assisted intervention–MICCAI 2012*, pp. 348–356. Springer, 2012.

[6] James Diamond, Neil H. Anderson, Peter H. Bartels, Rodolfo Montironi, and Peter W. Hamilton, "The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia," *Human Pathology*, vol. 35, no. 9, pp. 1121 – 1131, 2004.

[7] H. Chang, Y. Zhou, A. Borowsky, K. Barner, P. Spellman, and B. Parvin, "Stacked predictive sparse decomposition for classification of histology sections," *International Journal of Computer Vision*, pp. 1–16, 2014.

[8] J. C. Caicedo, A. Cruz, and F. A. Gonzalez, "Histopathology image classification using bag of features and kernel functions," in *Artificial Intelligence in Medicine*, Carlo Combi, Yuval Shahar, and Ameen Abu-Hanna, Eds., vol. 5651 of *LNCS*, pp. 126–135. Springer Berlin Heidelberg, 2009.

[9] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR, 2006 IEEE Computer Society Conference on*, 2006, vol. 2, pp. 2169–2178.

[10] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, June 2009, pp. 1794–1801.

[11] S. Petushi, F. U. Garcia, M. M. Haber, C. Katsinis, and A. Tozeren, "Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer," *BMC Medical Imaging*, vol. 6, no. 14, 2006.

[12] Y. Zhang, B. Zhang, F. Coenen, and W. Lu, "Breast cancer diagnosis from biopsy images with highly reliable random subspace classifier ensembles," *Machine Vision and Applications*, vol. 24, no. 7, pp. 1405–1420, 2013.

[13] Sonal Kothari, John H. Phan, Andrew N. Young, and May D. Wang, "Histological image classification using biologically interpretable shape-based features," *BMC Medical Imaging*, vol. 13, no. 1, pp. 1–17, 2013.

[14] H. Ahmady Phoulady, B. Chaudhury, D. Goldgof, L. O. Hall, P. R. Mouton, A. Hakam, and E. M. Siegel, "Experiments with large ensembles for segmentation and classification of cervical cancer biopsy images," in *SMC, 2014 IEEE International Conference on*, Oct 2014, pp. 870–875.

[15] P. S. Liao, T. S. Chen, and P. C. Chung, "A fast algorithm for multilevel thresholding," *Journal of Information Science and Engineering*, vol. 17, pp. 713–727, 2001.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR 2005, IEEE Computer Society Conference on*, June 2005, vol. 1, pp. 886–893 vol. 1.

[17] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627–1645, Sept 2010.