Cervical cancer histology image identification method based on texture and lesion area features

Lisheng Wei\textsuperscript{a}, Quan Gan\textsuperscript{b} and Tao Ji\textsuperscript{b}

\textsuperscript{a}Anhui Key Laboratory of Detection Technology and Energy Saving Devices, Anhui Polytechnic University, Wuhu, China; \textsuperscript{b}School of Electrical Engineering, Anhui Polytechnic University, Wuhu, China

**ABSTRACT**

The issue of an automated approach for detecting cervical cancer is proposed to improve the accuracy of recognition. Firstly, the cervical cancer histology source images are needed to use image preprocessing for reducing the impact brought by noise of images as well as the impact on subsequent precise feature extraction brought by irrelevant background. Secondly, the images are grouped into ten vertical images and the information of texture feature is extracted by Grey Level Co-occurrence Matrix (GLCM). GLCM is an effective tool to analyze the features of texture. The textures of different diseases in the source image of Cervical Cancer Histology (such as contrast, correlation, entropy, uniformity and energy, etc.) can all be obtained in this way. Thirdly, the image is segmented by using K-means clustering and Marker-controlled watershed Algorithm. And each vertical image is divided into three layers to calculate the areas of different layers. Based on GLCM and lesion area features, the tissues are investigated with segmentation by using Support Vector Machine (SVM) method. Finally, the results show that it is effective and feasible to recognize cervical cancer by automated approach and verified by experiment.

**KEYWORDS**

Cervical cancer; automated approach; segmentation; GLCM; SVM

1. Introduction

Cervical cancer is the second leading cause of mortality in women worldwide \cite{1}. There are varieties of cervical cancers, of which Squamous Cell Carcinoma and Adenocarcinoma are two main types. It is reported by Ministry of Health (MOH) of the People’s Republic of China that everyday there were over 100,000 new cases of cervical cancer in 2016, more than a third of which died. And the vast majority of deaths take place in developing countries without much access to health care. The incidence and mortality of cervical cancer in women is very high in Central and Western China. And Human papillomavirus (HPV) is one of the most common sexually transmitted infections in the world. Papanicolaou (Pap) smear is a traditional method for detecting abnormal cervical cell. However, Pap test for cervical detection suffers from variability and no specificity, which may be caused by inaccurate detection in the hospital.

Cervical intraepithelial neoplasia (CIN) can be divided into three categories: low-grade lesion (CINI), high-grade lesion (CINII or CINIII) and carcinoma. Epithelial tissue can be sub-divided into three layers: the basal layer, the intermediate layer and the superficial layer. In CINI, the lower third of the epithelial tissue could be formed by proliferative lesions. In CINII, the upper two thirds of the epithelial tissue could be brought about by proliferative lesions. In CINIII, the whole epithelium is proliferative lesions. So an effective approach capable of analyzing and identifying the lesion would be a good option. Image processing technologies have enjoyed increasingly wide application to clinical practice, such as Computed Tomography (CT), Digital Subtraction Angiography (DSA), and Magnetic Resonance Imaging (MRI). The mathematical theory of image processing has been extensively explored for cervical cancer diagnosis. Peng Guo et al. researched Nuclei-Based features for uterine cervical cancer histology image analysis with fusion based classification. They developed new features for CIN classification of segment epithelium regions and including nuclei ratio, a cellular area, combination and layer-by-layer triangle. Payel Rudra Paul et al. introduced an automated cervical detection method based on the Pap smear images \cite{2}. Yinhai Wang et al. presented an automated computer-assisted system for the diagnosis of cervical...
intraepithelial neoplasia (CIN) using ultra-large cervical histological digital slides [3]. However, the method is time-consuming and too slow to be used as a clinical medicine device. Soumya De et al. investigated an automated, located, fusion-based approach for cervix histology image analysis for CIN classification. Results showed that at least 15.5 CIN exact grade classification improvement was applied in vertical segment method [4]. Gisele Helena Barboni Miranda et al. discussed the proposal, the implementation and the evaluation of a methodology for the analysis of CIN images. Also, clustering algorithms, graph morphology and Delaunay Triangulation were applied to obtain the histological layers of the epithelial tissue [5]. Stephen J. Keenan et al. introduced an automated machine vision system for the histological grading of CIN [6]. The features of cervical squamous epithelium were analyzed by employing iterative threshold, and the images were divided into three layers based on the Delaunay Triangulation algorithm. The result showed better accuracy in normal and CIN III proliferative lesions, yet not good in CINII proliferative lesions. Miranda GHB et al. [7] introduced computational techniques in the process of histopathological images using a watershed segmentation method, Neighborhood Graphs theory and Complex Networks were evaluated concerning the detection of the presence of lesions in the tissue. The result displayed that the maximum accuracy obtained in the evaluation of the detection of abnormalities was 88%. Wang Y et al. [8] presented a color-based approach for automated segmentation in tumor tissue classification. The method includes color normalization, elimination of tedious and time-consuming steps and principal component analysis. Experimental studies showed that it could be applied to other microscopic images prepared with the same type of tissue staining. M. Guillaud et al. [9] used the colposcopically directed and histopathologically classified cervical biopsy as the gold standard. The agreement of a test and training set showed that the sets created randomly were indeed similar, and that the discrimination score worked equally well in both sets of cells, thus proving that this study would be more useful. M. Veta P et al. [10] introduced fast digital slide scanners that provide whole slide images. It includes: pre-processing with color unmixing and morphological operators, marker-controlled watershed segmentation at multiple scales and with different markers, post-processing for rejection of false regions, merging of the results from multiple scales.

In this paper, the automatic analysis together with the diagnosis of Cervical Cancer using image processing technology and feature extraction method are described. The rest of the paper is organized as follows: In Section II, the identification process of Cervical Cancer is presented, including (A) image processing, (B) the preprocessing of Texture Feature Extraction and (C) the preprocessing of Lesion Area Feature Extraction. In part B, the experimental results and analysis using image vertical segmentation and texture feature extraction are presented. The part C presents the experimental results and analysis using K-means clustering and Marker-controlled watershed segmentation, image vertical segmentation and lesion area feature extraction of three layers. Two comparison test results and analyses are presented in Section III. One shows that the results of cervical cancer are identified by GLCM, Lesion Area and GLCM + Lesion Area method, respectively. The other shows that the results of different methods are identified by the method in Reference [4], the method in Reference [6] and the method GLCM + Lesion Area features proposed in this paper. Section IV shows the conclusion.

2. Identification method of cervical cancer

The image analysis techniques contain five phases: (1) the vertical segment of the squamous epithelium; (2) texture feature extraction; (3) K-means clustering and Marker-controlled watershed segmentation; (4) lesion areas feature extraction of three layers; (5) the experimental results of SVM recognition using GLCM + Lesion Area method.

Figure 1 shows the block diagram of the whole image approach and the calculation of proliferative lesions areas. In Figure 1, the concrete method of aiding diagnosis of cervical cancer can be described as follows: (A) image preprocessing; (B) the preprocessing of texture feature extraction; (C) the preprocessing of lesion areas feature extraction.

A. Image preprocessing

In this paper, the cervical intraepithelial neoplasia (CIN) images were obtained from Bengbu Polytechnic Institute, which provided materials from cervical uterine histological exams. Figure 2 shows the CIN grading image samples. The images analyzed include 90(normal30, CIN lesions 30, cancer 30).

Firstly: To simplify the experiment, the part of image $I_c$ from Figure 3(a) was chosen. Image samples are often degraded by noise during the formation, transmission, and recording processes [11–13]. Hence, the method of median filter is used for image denoising so that the images can be obtained smoothly. The median filter procedure contains four steps: (1) filter
template is used in image based on virtual navigation and template center should be overlapped with pixel; (2) the gray value can be calculated in the image; (3) sorting the gray value from lowest to highest; (4) storing intermediate data to the center of image pixels.

Secondly: In order to acquire a better Euclidean distance transform image, Ic is rotated to an angle $\theta$. We used the nearest neighbor method to process. When the image is rotated an angle, the width and height of new image can be displayed as follows: where, $width_{new}$ and $height_{new}$ are the width and height of new image, $width$ and $height$ are the width and height of original image.

$$
width_{new} = 2 \times \left( \frac{width}{2\times\cos\theta} + \frac{height}{2\times\sin\theta} \right) \\
height_{new} = 2 \times \left( \frac{height}{2\times\cos\theta} + \frac{width}{2\times\sin\theta} \right)
$$

(1)

where, $width_{new}$ and $height_{new}$ are the width and height of new image, $width$ and $height$ are the width and height of original image.

And the rotation formula is shown as follows:

$$
\begin{align*}
(x_0 - x_1) &= (x_1 - x_2) \times \cos\theta - (y_1 - y_2) \times \sin\theta \\
(y_0 - y_1) &= (x_1 - x_2) \times \sin\theta + (y_1 - y_2) \times \cos\theta
\end{align*}
$$

(2)

where, $x_0$ and $y_0$ are the original pixels; $x_1$ and $y_1$ are the new pixels; $x_1$ and $y_1$ are the two rotating center coordinates of original image; $x_2$ and $y_2$ are the two rotating center coordinates of new image.

The image $I_c$ is rotated $\theta^\circ$ ($\theta = 30^\circ$) and reached horizontal coordinate, as shown in Figure 3(c).

Thirdly: Euclidean distance transform is widely used in binary image, especially in skeleton extraction and interpolation. In the paper, the histology image should be operated by Euclidean distance transform to determine the highest intensity pixel. In Euclidean distance transform, a binary image consists of target pixel and background pixel. The result of Euclidean distance transform is a gray level texture image (range image). A 2D binary $N \times N$ image array can be represented by $a_{ij} = 0$ or $1 (i,j = 0,\ldots,N-1)$, so the binary image can be described as $B = \{ (x,y) : a_{x,y} = 1 \}$ [14]. The Euclidean distance of pixel $a_{ij}$ is computed by:

$$
d_{ij}^2 = \min_{(x,y) \in B} (i-x)^2 + (j-y)^2
$$

(3)

In order to obtain the highest intensity pixel, the image should be used in binary process. In Figure 4(a) the binary mask image $I_c^B$ should be created. Figure 4(b) shows the Euclidean distance transform image $I_c^E$. Numerous highest intensity pixels are presented in Figure 4(b). To obtain the vertical segment image, the medial axis is kept in $I_c^E$.

B. The preprocessing of texture feature extraction

1) Image vertical segmentation

Image analysis is arousing increasing interest in mathematics. In particular, in medicine, many new tomographic imaging modalities are now producing
different images of organs [15–17]. The measurement of precancerous changes in squamous epithelium [3] can be simplified by medial axis transformation.

A four-step approach is used for the vertical segment images:

**Step 1:** The bounding rectangle box $r_{box}$ is created to restrict the image region and it is presented in the orange dotted line. Then the highest intensity pixel can be obtained based on Euclidean distance transformation shown in Figure 3(b).

**Step 2:** The edge image of the epithelium point should be determined. In $R^2$, the skeleton is a set of curves centered on 2D-space. In $R^3$, the skeleton is a set of surfaces and curves in 3D-space. On the skeleton, there existing points which can be classified into four types: boundary point, branching point, regular point, and arc point [14]. The four types of points are shown in Figure 5.

**Step 3:** In Figure 4(b), the highest intensity pixel $l_{middle}$ can be described in Figure 6(a), but it is unnecessary to draw the leftmost and rightmost two lines. To be more specific, if the intersection of the leftmost and rightmost edge is composed of three or more points, it is the regular point. If the edge is composed of one point, it is the boundary point. If the edge is composed of intersecting lines then the intersection

![Figure 3](image1.png)

**Figure 3.** (a) Original image, (b) Rotated original image, (c) Part of image from (b).

![Figure 4](image2.png)

**Figure 4.** Euclidean distances transform. (a) The binary image $I_{cB}$, (b) Distance transform image $I_{cE}$.

![Figure 5](image3.png)

**Figure 5.** The four types of points. (a) Regular point, (b) Branching point, (c) Boundary point, (d) Arc point.

![Figure 6](image4.png)

**Figure 6.** Create medial axis. (a) The highest intensity pixel, (b) line transform based on step 2 and step 3.
point is the branching point. If the edge is similar to the segment of the circle or ellipse and the intersection point is the arc point. So the leftmost and rightmost two lines can be transformed by using the four types of points. In step 2, it is found that the leftmost edge and the rightmost edge are similar to Figure 4(a) (Regular point). The leftmost and the rightmost two lines can be merged into one line (based on Pappus Law) and the final points are determined by midpoints of proliferative lesions, as shown in Figure 6(b) (the light blue arrow). If the points are similar to Figure 5(b), (c), there is no need to process the leftmost and rightmost edge like regular point, and Distance transform can determine the leftmost and rightmost lines directly. If the points are similar to Figure 5(d) (arc point), the furthest point can be regarded as the leftmost and rightmost edge points. The distance-transform improvement of [4] had difficulties finding left and right axis in nearly rectangular and triangular regions [1]. So the incorrect medial axis estimation from [4] needs to be overcome [4] by using line transform based on the four types of points. The result is shown in Figure 6(b).

Step 4: The perpendicular line for each point on the main axis can be determined. So the epithelium can be divided into ten vertical image regions, as done in [4]. Figure 7 shows the ten vertical images regions.

After the ten vertical segment images are determined, we will deal with these local images to recognize cervical cancer.

2) Texture feature extraction

Texture feature extraction is usually proposed with the help of statistical, spectral or gray level co-occurrence matrix (GLCM) methods [19–22]. Gray level co-occurrence matrix is an effective method to analyze texture features. The information of direction, interval and changing range can be reflected by GLCM. The image retrieval is presented as follows. Firstly, the image should be processed with Gray scale quantization; Secondly, it is necessary to construct gray level co-occurrence matrix from four directions: horizontal direction, vertical direction, diagonal direction and back diagonal direction. Thirdly, constructing eigenvectors and feature matrix should be transformed by Gaussian normalization. Fourthly, the eigenvectors should be calculated based on the three steps above. Lastly, the image should be matched with the eigenvectors of feature matrix. On CIN grade image, the GLCM method can perform better than statistical and spectral methods. The texture features used in the study include five features which are Contrast, Correlation, Entropy, Homogeneity, and Angular Second Moment.

Figure 7. Ten vertical image regions ($L_1, L_2, \ldots, L_{10}$).
The GLCM matrix describes the texture of an image by measuring how often do pairs of pixel have specific values with a specified relationship of spatial pixels occurred in an image [19]. GLCM is calculated based on distance and angle. Throughout the paper, four directions are examined \( \{ \theta = 0, 45, 90, 135 \} \) with three distances (1, 2, and 3 pixels). \( G(i,j) \) is the gray level co-occurrence matrix, \( L \) is gray level, \( i \) and \( j \) are gray level pixels.

(1) Contrast

\[
A_1 = \sum_{i} \sum_{j} (i-j)^2 G(i,j)
\]  

(4)

In Equation 4, \( |i-j| \) refers to the gray level difference between adjacent pixels. \( G(i,j) \) is the distribution probability of gray level difference between adjacent pixels. The Contrast is mainly applied to describe the degree of depth of the image textile grooves. The contrast feature is a difference moment of the regional co-occurrence matrix and is a measure of the contrast or the amount of local variations present in an image [18].

(2) Correlation

\[
A_2 = \frac{\sum_{i} \sum_{j} (i - \bar{x})(j - \bar{y}) G(i,j)}{\sigma_x \sigma_y}
\]  

(5)

In Equation 5, \( A_2 \) refers to correlation, \( \bar{x} \) the mean value of the sum of the elements in each column in the square, the mean value of the sum of the elements in each row, \( \sigma_x \) the standard deviation of the sum of the elements in each column, and \( \sigma_y \) the standard deviation of the sum of the elements in each row. The Correlation is mainly used to describe the possible elements of each row and column in the process of vertical image segmentation.

(3) Entropy

\[
A_3 = -\sum_{i} \sum_{j} G(i,j) \log G(i,j)
\]  

(6)

The Entropy is used to measure the quantity of information within the image. If the image is not texturally uniform, the value is high. Otherwise, the value is low.

(4) Homogeneity

\[
A_4 = \sum_{i} \sum_{j} \frac{1}{(i-j)^2 + 1} G(i,j)
\]  

(7)

Homogeneity is used to measure the closeness of the distribution of pixels in the diagonal segmentation.

(5) Angular Second Moment

\[
A_1 = \sum_{i} \sum_{j} G^2(i,j)
\]  

Angular Second Moment is applied to describe the sum of squared elements in horizontal and vertical directions in the GLCM.

Five texture features were calculated, and the results of texture features are shown in Figures 8–12. Figures 8–12 each shows information of the five texture features: Contrast, Correlation, Entropy, Homogeneity, and Angular Second Moment. Contrast is illustrated in Figure 8, from which we can conclude that CIN lesion has the highest contrast values, distinguishing greatly CIN lesion from normal (or cancer). However, the Contrast feature of normal and cancer may be hard to differentiate. In Figure 9, it is shown...
that the Correlation of cancer has the lowest values so it can be distinguished from others. However, it is hard to recognize the Correlation of CIN and normal one. The Figures 10–11 show that the Entropy and Homogeneity features are somewhat higher than cancer tissues. The Entropy of the normal tissue is lower than the CIN lesion tissue in Figure 10. The Angular Second Moment characteristic of the texture is shown in Figure 12, which is the non-uniform and complexity of the curves. From Figures 8–12, it can be concluded that the five texture features based on GLCM can distinguish three different types of tissues. So it is effective to recognize the three different layers by local and texture feature method.

C. The segmentation of cells of lesion area feature extraction

1) The segmentation of k-means clustering and marker-controlled watershed

Image segmentation is an important method to extract image regions, and can be mainly categorized into three tasks. The first task is to identify and enhance edges, the second to group the edge pixels, and the third to assign a semantic label to each pixel. Cervical cell will undoubtedly, more or less, miss much information in the results of image segmentation. No matter what segmentation algorithm is taken, the accuracy of recognition will be affected. Thus, the marker-controlled watershed algorithm and clustering algorithm can reach a better segmentation in cervical cancer histology image. A major reason for this is that the cell can be in an elliptical or circular shape. Through the image samples and the features of nuclei, the watershed algorithm can divide the squamous epithelium image into different small regions. A previous study [21–23] also exploited the watershed method that has been widely used in medical image segmentation. Watershed segment method is based on topological theory. The grey scale image is modeled as a topographical model of a terrain, where the pixel value denotes the altitude of that point [24]. The method includes two steps: ranking process and submerging process. The gray scale pixel should be sorted from highest to lowest, and local minimum should be judged and marked by First Input First Output structure from lowest to highest as well. However, the result shows that watershed algorithm method is unable to restrain the over-segmentation of image effectively. Hence, the method should be improved to overcome the over-segmentation. A marker-controlled watershed algorithm was developed to locate breast tumor, and the data shows the detection rate is 90%.
Previous work [5] used watershed from markers to segment squamous epithelium image. Figure 13 shows the segmentation results by selecting different thresholds.

From Figure 13, we can find that different thresholds display different results. When 15, 45 and 80 (Figure 13(a), (c), and (d)) are selected by thresholds, image segmentation is not effective enough. Figure 13(b) shows the result better, still with some parts unmarked. The clustering algorithm is used to make up the deficiency above. From squamous epithelium image it is concluded that the proliferative lesions mainly appear in nuclei areas. It is known that the cells are in the shape of circle or ellipse, and K-means clustering algorithm is suitable for nuclei segmentation. Clustering is an unsupervised learning algorithm and the method is usually very productive [25–28]. The description of the method is as follows:

**Step 1:** Give an initial set of \( K : O_1(c), O_2(c), \ldots, O_K(c) \). They each stand for the proliferative lesions, normal regions, and the other regions that are marked based on similarity. Parameter is used to count the loops.

**Step 2:** The vector set \( \{x\} \) is used as a subset based on step 1, the subset can be described as follows: \( T_1(n), T_2(n), \ldots, T_K(n) \). Where \( K \) is the index of the \( K^{th} \) input.

**Step 3:** The Euclidean distances between the normalized input vector \( x \) and the center of each cluster \( O_i(n) \) can be concluded as: \( d = ||x - O_i(n)|| \), where \( d \) is the Euclidean distances.

**Step 4:** The center of new cluster can be calculated by using:

\[
O_c(n + 1) = \left(1/N_c\right) \sum x(x \in T_c(n))
\]

where \( O_c(n + 1) \) is the center of the \( c^{th} \) cluster consisting of \( N_c \).

**Step 5:** If the center of new cluster equates to \( O_c(n + 1) = O_c(n) \), the operation should repeat step 2.

For this work, K-means clustering is proposed to automatically define markers as an input with Marker-controlled watershed method to segment the regions that are not marked before [29]. So the regions are processed by using the K-means clustering and Marker-controlled watershed to improve the effectiveness of segmentation. Figure 14 shows the result of image segmentation. Thus, it can be concluded that the normal cell and cancer cell can be segmented well. The normal cell is small. Compared with normal cell, the CIN lesion cell is bigger and the cancer cell is the biggest one among them three. The main cause of the phenomenon is that the nuclei become less viscous. So the Marker-controlled watershed and clustering segmentation algorithm will keep a large amount of information in histology images.

### 2) Lesion area feature extraction

When the squamous epithelium image is segmented, we also need to obtain the ten vertical images regions. In order to display the result of image segmentation better, the Figure 15(a) should be processed by morphological operation, which, based on the set theory, can extract object features by suitably shaped structuring elements. It is also widely used in the edge of two-dimensional images, skeletons of objects, and convex hulls. Besides, it can still be employed for dilation, opening and closing, holes filling, and erosion.
So Morphological operation is an important way in image processing. In this paper, the nuclei of squamous epithelium may have bad effects (mainly segmented nuclei) after segmentation; in Figure 15(a), the image (orange areas) should be processed with Morphological construction. The image $I_0$ can be constructed by using holes filling. For example, we define the Figure 15(a) as binary image $I$, the mark images $F$, and the image of edge detection value is $1-I$:

$$
F(x, y) = \begin{cases} 
1 - I(x, y), & (x, y) \text{ is the edge of } I \\
0, & \text{if other}
\end{cases} \quad (10)
$$

$H$ is the transformed image, and we define $R_1(F)$ as $F$ constructed $I$.

The results of image segmentation are shown in Figure 16(a–j).

Ten segment images were obtained by image segmentation. Each segment image has a top layer, a middle layer and basal layer. So there are altogether 10 top layers, 10 middle layers and 10 basal layers. Through Figure 16 the segment images can be divided into three layers. As is shown in Figure 17, the first is basal layer, the second middle layer, and the last top layer. Thus the nuclei areas of different layers can be counted. Tables 1–3 show the areas of the squamous epithelium tissues and the curves are shown in Figures 18–20.

In Figure 18, it is apparent that the area of top layer of cancer curve is the highest. The average of the area of top layer of cancer curve is 623.3. However, the area of top layer of normal tissue is the lowest. The average of the area of top layer of normal tissue is 147.4. The area of top layer of CIN lesion tissue is lower than cancer areas. The area of top layer of normal is lower than CIN lesion tissue. The average of the area of top layer of lesion tissue is 199.3. From the Figure 19, we find that the areas of middle layers of cancer curve are the highest. The average of the areas of middle layers of cancer curve is 537.4. While the
areas of middle layers of normal tissue are the lowest. The average of the areas of top layers of normal tissue is 200.8. The areas of middle layers of CIN lesion tissue are lower than cancer areas. The average of the areas of middle layers of lesion tissue is 262.9. And it is clear that the areas of basal layers of cancer curve are the highest from the Figure 20. The average of the areas of basal layers of cancer curve is 525.3. But the areas of basal layers of normal tissue are the lowest. The average of the areas of basal layers of normal tissue is 123.6. The areas of basal layers of CIN lesion tissue are lower than cancer areas. The areas of basal layers of normal are lower than CIN lesion tissue. The average of the areas of middle layers of lesion tissue is 167. Then we get the results of the normal layers areas, CIN lesion layers areas and cancer layers areas in Tables 1–3 respectively.

From Tables 1–3 and Figures 18–20, several conclusions can be drawn. First, the areas of normal and CIN lesion tissues are lower than cancer areas. Moreover
the nucleus of normal cell is small and regular. Second, the floating ranges of cancer areas are the highest, and the ratio of nuclei area is increasing with the proliferative lesion. Third, Table 1 and Figures 18–20 present that the curve of normal layers areas is flat, a major cause of which is normal cervical cells.

Forthly, the area of basal, middle, top layers are different. In normal layers, the area of middle layers is higher than the areas of basal and top layers. In CIN lesion layers, the area of the three layers is irregular. In cancer layers, the area of the certain layers may be very high. It is also displayed in Table 3 that there are high values in the basal and top layers. In basal layers, the values of areas are 741,775, 659,618,814,599. In middle layers, the value of area is 520. In top layers, the values of areas are 686, 649, 646, 1104, and 1138.

3. Experimental results

Support Vector Machine (SVM) is derived based on statistical learning theory. SVM is suitably used as learning classifier of small size, and characterized by overall optimum, strong generalization ability. The other classification, such as artificial neural networks, can not avoid being trapped in local optimum and the reorganization of result could have biased. One of the biggest advantages of SVM is the ability to choose kernel function in high-dimensional space. SVM also has good robustness. SVM is performed by creating the n-dimensional hyperplane that separates data into two classes. The goal of SVM is to pinpoint an optimum hyperplane which can separate clusters from vectors to categorize the target variables on one side and the other class on the other side. The support vectors refer to the vectors near the hyperplane.

To extract the parameter of feature and area, we obtain the classifications by SVM [30–31]. When the extracted texture features including correlation, entropy, uniformity and energy, and the feature of the lesion area are taken as the inputs of SVM, some requirements are needed for the penalty factors in training samples, because the bigger the penalty factor is, the less the error can be tolerated.

![Figure 19. The area of middle layer.](image)

![Figure 20. The area of basal layer.](image)

| Vertical image | Basal area | Middle area | Top area | Basal area | Middle area | Top area |
|----------------|------------|-------------|----------|------------|-------------|----------|
| 1  | 113        | 232         | 192      | 6  | 178        | 193       | 142      |
| 2  | 70         | 197         | 175      | 7  | 172        | 193       | 155      |
| 3  | 123        | 233         | 64       | 8  | 114        | 180       | 127      |
| 4  | 69         | 192         | 137      | 9  | 182        | 187       | 141      |
| 5  | 96         | 210         | 124      | 10 | 119        | 191       | 217      |

Table 1. The result of normal layers areas.

| Vertical image | Basal area | Middle area | Top area | Vertical image | Basal area | Middle area | Top area |
|----------------|------------|-------------|----------|----------------|------------|-------------|----------|
| 1  | 222        | 230         | 114      | 6  | 200        | 144       | 120      |
| 2  | 90         | 278         | 159      | 7  | 207        | 143       | 268      |
| 3  | 263        | 520         | 229      | 8  | 73         | 203       | 333      |
| 4  | 91         | 175         | 94       | 9  | 178        | 305       | 193      |
| 5  | 178        | 292         | 240      | 10 | 168        | 339       | 243      |
In this paper, three kinds of diseases known as normal, CIN lesion and cancer are identified by Texture feature and Lesion area feature. Three kinds of diseases are represented as Class I, Class II and Class III, respectively. Texture feature classifier is the SVM1, Lesion area feature classifier refers to SVM2, and integrated classifier is represented as SVM. Set the classification of the output results of SVM1 as weight classification of the output results of SVM2 as classification. In accordance with the proportion of $T_1$ and $T_2$, the diseases types can be output.

The 30 images (there are 20 testing data set and ten standard data set). We define training sample as $(x_i, y_i), i=1, 2, \cdots, l, x \in R^n, y \in \{\pm 1\}$, and hyper plane is $(\omega \cdot x) + b = 0$. The formulas are as follows:

$$\Phi(\omega) = \frac{1}{2} ||\omega||^2, y_i[(\omega \cdot x) + b] \geq 1, i=1, 2, \cdots, l$$ \hspace{3.5cm} (12)

There are three types of kernel functions: the Gaussian RBF kernel function, polynomial kernel function and linear kernel function.

1) Gaussian RBF kernel function:

$$K(x'_i, x'_j) = \exp\left(-\frac{||x'_i - x'_j||^2}{2\sigma^2}\right)$$ \hspace{3.5cm} (13)

2) Polynomial kernel function:

$$K(x'_i, x'_j) = [(x'_i \cdot x'_j) + 1]^q$$ \hspace{3.5cm} (14)

3) Linear kernel function:

$$K(x'_i, x'_j) = x'_i \cdot x'_j$$ \hspace{3.5cm} (15)

where $x'_i$ is the predictions parameter; $\sigma$ and $q$ are parameters of kernel function. The three kernel functions all have a restricted range. Gaussian RBF kernel function tends to localize characteristics, and it is suitable for closer-in data.

In order to classifying nonlinear data, the kernel function $K(x'_i, x'_j)$ is introduced and made linearly separable in high dimensional space, shown as $K(x'_i, x'_j) = \langle \phi(x'_i), \phi(x'_j) \rangle$. This paper picks up Gaussian kernel function or Radial Basis Function, biasing towards local characteristics, which is generally applicable to data points with a relatively closer distance. Nevertheless, in this paper, the image is divided into 10 areas through vertical image segmentation and the local characteristics of the 10 areas are also analyzed through feature extraction of texture and lesion areas. The Gaussian kernel function in the classifier of Support Vector Machine is as follows:

$$K(x'_i, x'_j) = \exp\left(-\frac{||x'_i - x'_j||^2}{2\sigma^2}\right)$$ \hspace{3.5cm} (16)

where $\sigma$ is a parameter value in the Radial Basis Function.

The optimal function:

$$J(x) = \text{sgn}\left(\sum_{i=1}^{l} y_i x_i K(x'_i, x'_j) + b\right)$$ \hspace{3.5cm} (17)

where, sgn is sign function; $l$ is sample; $x'_i$ is the number of training sample; $i$ and $j$ are any of the samples from 1 to $l$; $y_i$ is the output of training sample; $x_i$ is the LaGrange coefficient; $b$ is the offset.

Table 4 shows the results of cervical cancer identified by GLCM, Lesion Area and GLCM + Lesion Area methods, respectively.

Table 5 shows the results of different methods identified by the methods in reference [4], the method in reference [6] and the method GLCM + Lesion Area features proposed in this paper.

Several conclusions can be drawn from Table 5. (1) From normal recognition, 18 images can be recognized by the method in reference [4], and 14 images by the method in reference [6]. The recognition rate of GLCM and Lesion Area features proposed in this paper is equivalent to that of reference [4]. And the Current classification can get 90 percent. (2) For CIN lesion recognition, only 12 images can be recognized by the method in reference [4], and 14 images by the method in reference [6]. However, only 14 images can be recognized by the results of proposed method in this paper. (3) For cancer recognition, 16 images can be recognized by the method in reference [4], and 15

---

Table 3. The result of cancer layers areas.

| Vertical image | Basal area | Middle area | Top area | Vertical image | Basal area | Middle area | Top area |
|----------------|------------|-------------|----------|----------------|------------|-------------|----------|
| 1              | 235        | 230         | 164      | 6              | 775        | 144         | 474      |
| 2              | 150        | 278         | 686      | 7              | 659        | 143         | 646      |
| 3              | 249        | 520         | 450      | 8              | 618        | 203         | 517      |
| 4              | 413        | 175         | 649      | 9              | 814        | 305         | 1104     |
| 5              | 741        | 292         | 405      | 10             | 599        | 339         | 1138     |
images by the method in reference [6]. However, 17 images can be recognized by the results of proposed method in this paper. And the Current classification can get 90 percent. (4) Two previous works are presented as contrast experiments by adopting the same method. The result shows that the recognition of CIN lesion is worse than that of previous work, but the recognition of normal and cancer tissues is better than that of previous work. So the results show that it is effective and feasible to recognize cervical cancer by current experiment.

4. Conclusion

In this paper, a localized and automated analysis of cervical histological images is presented to identify the CIN degree. New features include GLCM features (Contrast, Correlation, Entropy, Homogeneity, and Angular Second Moment), and the area features. (1) Vertical image segmentation: four types of points are utilized (regular point, branching point, boundary point, arc point) to determine the leftmost and rightmost edges, and the leftmost and rightmost edges can be transformed exactly. (2) GLCM features: five parameters are extracted to recognize three different layers. The results show that the texture feature can reflect the cervical cancer as a whole from Figures 8–12. (3) Image segmentation: Marker-controlled watershed and clustering algorithm are applied to segment images. The segmentation algorithm may help us to achieve the desired result based on elliptical shape or circular shape of the cell. (4) Area features: the segmented image is divided into three layers and calculated the areas, and the result shows that cancer tissue can reach a better recognition than the others. Finally, SVM is used to classify data, and the result shows that only 12 to 15 images can be recognized from GLCM feature and Lesion Area feature, but the GLCM + Lesion Area features can recognize 18 images. Based on these, current work can get a higher accuracy (90 percent) than previous work. However, the sample size is too small to achieve the conclusion of the proposed method. How to increase sample data, reduce conservation and make the results general applicability is one of the most important issues to be investigated in the future.

Acknowledgement

All authors declare that there is no conflict of interest regarding the publication of this paper. And the authors would like to thank anonymous reviewers and the editor for their valuable comments and helpful suggestions. And the authors would like to thank Prof. Fei for his supports that have improved the quality of the paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Natural Science Foundation of Anhui Province under grant 1608085MF146, the Natural Science Research Program of Colleges and Universities of Anhui Province under grant KJ2016A062, the Visiting Study Foundation for Outstanding Young Talent of Anhui Educational Committee under grant gxfxZD2016108, and the Foundation for talented young people of Anhui Polytechnic University under grant 2016BJRC008.

References

[1] Guo P, Banerjee K, Joe Stanley R, et al. Nuclei-based features for uterine cervical cancer histology image analysis with fusion-based classification. IEEE J Biomed Health Inform. 2016;20:1595–1607.
[2] Payel Rudra P, Mrinal Kanti B, Debotosh B. Automated cervical cancer using pap smear images. Adv Intell Syst Comput. 2015;335:267–278.
[3] Yin-Hai W, Crookes D, Eldin OS, et al. Assisted diagnosis of cervical intraepithelial neoplasia (CIN). IEEE J Sel Top Signal Process. 2009;3:112–121.
[4] De S, Joe Stanley R, Lu C, et al. A fusion-based approach for uterine cervical cancer histology image classification. Comput Med Imaging Graph. 2013;37:475–487.
[5] Gisele Helena Barboni M, Junior B, Edson Garcia S, et al. Structural Analysis of Histological Images to Aid Diagnosis of Cervical Cancer. 25th SIBGRAPI Conference on Graphics, Patterns and Images. 2012 August 22-25; Ouro Preto, Brazil.
[6] Keenan SJ, Diamond J, McCluggage WG, et al. An Automated Machine Vision System for The Histological Grading of Cervical Intraepithelial Neoplasia (CIN). J Pathol. 2000;192:351–362.
[7] Miranda GHB, Soares EG, Barrera J, et al. Method to Support Diagnosis of Cervical Intraepithelial Neoplasia
(CIN) Based on Structural Analysis of Histological Images. Proceedings of the IEEE Symposium on Computer-Based Medical Systems (CBMS); 2012. p. 1–6.

[8] Yi-Ying W, S-C, Chang L-W, Wu, et al. A Color-Based Approach for Automated Segmentation in Tumor Tissue Classification. Proceedings of The 29th Annual International Conference of The IEEE EMBS. 2007 August 23-26; Lyon, France.

[9] Guillaud M, Adler-Storthz K, Malpica A, et al. Subvisual chromatin changes in cervical epithelium measured by texture image analysis and correlated with HPV. Gynecol Oncol. 2006;99:16–23.

[10] Veta M, van Diest PJ, Kornegoor R, et al. Automatic nuclei segmentation in H&E stained breast cancer histopathology images. PLoS One. 2013;8:e70221.

[11] Jalal Fadili M, Starck J-L, Bobin J. Image decomposition and separation using sparse representations: an overview. Proc IEEE. 2010;98:983–994.

[12] Li-Wei K, Chia-Wen L, Yu-Hsiang F. Automatic single-image-based rain streaks removal via image decomposition. IEEE Trans Image Process. 2012;21:1742–1755.

[13] Panetta K, Bao L, Agaian S. Sequence-to-sequence similarity-based filter for image denoising. IEEE Sensors J. 2016;16:4380–4388.

[14] Yu-Hua L, Shi-Jinn H, Seltzer J. Parallel computation of the euclidean distance transform on a three-dimensional image array. IEEE Trans Parallel Distrib Syst. 2003;14:203–212.

[15] Bonnassie F, Peyrin D, Attali A. A new method for analyzing local shape in three-dimensional images based on medial axis transformation. IEEE Trans Syst, Man, Cybern B. 2003;33:700–705.

[16] Lee DT. Medial axis transformation of a planar shape. IEEE Trans Pattern Anal Mach Intell. 1982;4:363–369.

[17] Wang S, Rosenfeld A, Wu AY. A medial axis transformation for grayscale pictures. IEEE Trans Pattern Anal Mach Intell. 1982;4:419–421.

[18] Jian Y, Jing-Feng G. Image Texture Feature Extraction Method Based on Regional Average Binary Gray Level Difference Co-occurrence Matrix. International Conference on Virtual Reality and Visualization (ICVRV). 2011 November 4-5; Beijing, China.

[19] Deepti Yadav M, Partha Sarathi, Malay Kishore Dutta. Classification of Glaucoma Based on Texture Features Using Neural Networks. Seventh International Conference on Contemporary Computing (IC3). 2014 August 7-9; Noida, India.

[20] Owen KK, Wong DW. An approach to differentiate informal settlements using spectral, texture, geomorphology and road accessibility Metrics. Appl Geography. 2013;38:107–118.

[21] Dragut L, Csillik O, Eisank C, et al. Automated parameterisation for multi-scale image segmentation on multiple layers. ISPRS J Photogramm Remote Sens. 2014;88:119–127.

[22] Nurhayati OD, Susanto A, Widodo TS, et al. Principal component analysis combined with first order statistical method for breast thermal images classification. Int Eng Technol Res J. 2011;2:72–78.

[23] Lewis SH, Dong A-J. Detection of Breast Tumor Candidates Using Marker-Controlled Watershed Segmentation and Morphological Analysis. IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI). 2012 April 22-24; Santa Fe, NM, USA.

[24] Sivagami M, Revathi T. Marker Controlled Watershed segmentation Using Bit-Plane Slicing. Int J Image Process Vision Sci. 2012;1:6–10.

[25] Peng-Fei S, Wen-Jian Q, Jie Y, et al. Segmenting Multiple Overlapping Nuclei in H&E Stained Breast Cancer Histopathology Images Based on An Improved Watershed. IET International Conference on Biomedical Image and Signal Processing (ICBISP 2015). 2015 November 19-19; Beijign, China.

[26] Tzortzis G, Likas A. The Global Kernel K-means Clustering Algorithm. IEEE World Congress on Computational Intelligence. 2008 June 1-8; Hong Kong, China.

[27] Rahmani M, Akbarizadeh G. Unsupervised feature learning based on sparse coding and spectral clustering for segmentation of synthetic aperture radar images. IET Comput Vision. 2015;9:629–638.

[28] Alush A, Goldberger J. Hierarchical image segmentation using correlation clustering. IEEE Trans Neural Netw Learning Syst. 2016;27:1358–1367.

[29] Lilla B, Silvia M, Alessia B, et al. Cluster Analysis Boosted Watershed Segmentation of Neurological Image. 4th International Congress on Image and Signal Processing (CISP). 2011 October 15-17; Shanghai, China.

[30] Chang C-C, Lin C-J. LibSVM: A library for support vector machines. ACM Trans Intell Syst Technol. 2011;2:1–27.

[31] Burges CJC. A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov. 1998;2:121–167.