Determining Optimal Textural Features for Cervical Cancer lesions using the Gaussian Function

S. Pradeep Kumar Kenny¹, Dr. S. Allwin²

¹Centre for Information Technology & Engineering, Manonmaniam Sundaranar University, Tirunelveli
²Dept. Of Computer Science and Engineering, Infant Jesus College of Engineering and Technology, Tuticorin, India.

Abstract— Cervical cancer is one of the most deadliest gynaecologic cancer occurring in women and it is also a very important research area in image processing. The main reason as to why this cancer is so deadly is that it cannot be detected early as it doesn’t throw any symptoms until it evolves to the final stages of cancer. The reasons attributed to this is the cancer itself and also to the lack of pathologists available to screen the cancer especially in third world countries. One solution would be to automate the system. But this comes with a set of problems as we have a number of features to determine whether the given cell is cancerous or not. If we calculate with too many features then the accuracy goes up but speed comes down. And if you reduce the features it is the other way round. To strike a balance between the speed and accuracy of a system in this paper we have proposed a unique ranking system using Gaussian function to determine the right balance of features which can be used in an automated system.

Keywords— Cancer, Cervical cancer, Cervical cytology Gaussian function, Textural features

I. INTRODUCTION

Cervical cancer is now considered as the contributor of most number of deaths among women worldwide. It is curable if it is detected and treated in its initial stage. But detecting the cancer in its initial stage is difficult because it shows symptoms only in the later stages. The traditional screening methods available like the visual procedures take a long time and often produce incorrect results. Further it is highly impossible for a few pairs of eyes to screen each and every woman on the planet. To solve this problem an automated process is needed. An automated process could not only accelerate the process but it also could give accurate results. Normally all automated systems consist of three phases [12] namely pre-processing phase for removing noise and image correction, segmentation phase to isolate the ROI form the rest of the image and feature extraction phase to extract the properties of the images using various mathematical formulas.

Image segmentation is an image analysis task wherein you individually decompose the various object present in an images. You can decompose the image into disjoint regions so that within each region the features will have strong, visual similarity, reasonable homogeneity and statistical correlation. Image segmentation algorithms are classified into different groups like region based techniques, feature thresholding, clustering techniques, contour based techniques, [1]-[11] etc.

All these approaches different advantages and limitations in terms of computational cost, performance, applicability and suitability.

Once segmentation is completed then the most important job is features extraction. In the real world we see a lot of things with our eyes. Although our eyes see the image it’s the brain that interprets that image. The brain describes each and every image using its shape, spatial location and textural properties. Color, brightness and contrast are all a subpart of textural properties. Using textural properties we can make out what a surface it really is. You can tell whether an object surface is rough, smooth, broken etc. A red ball and red apple from a distance can be distinguished by the texture of the object. Technically a texture is the measure of light refracted from the surface. A rough surface appears rough because half of the light bouncing out of the surface doesn’t bounce back the same way as the other half does. So if you take an old wooden plank then you see dark spots. It means the light which fell on that surface has bounced back the other way and hence your brain interrupts that as a rough surface.

There are plenty of mathematical formulas to determine these variances and that is the problem which is faced by an automated system. When you have too many features for detecting you are hard pressed as to which feature extraction technique needs to be used. If you use too much features it slows down the system and if you use too little then it brings down the accuracy. Hence a right balance has to be struck. Therefore the available features have to be ranked in some way so that we can have the right combination of features.

Nowadays with rise of new innovative computational methods there are plenty of ways to do this. In this paper we have proposed a novel ranking system by making use of Gaussian function. This paper has been organized into IV sections. Section II discusses how the features extracted and classified. And Section III shows the results. Conclusions are made in Section IV.

II. IMPLEMENTATION

The Block Diagram for this approach is given in Figure 1.
A. Input Image

We obtain the image of the cervix by extracting cells from the cervix using a swab and it is placed in a slide. This is then digitalised and the resulting image is obtained. A sample cervical image is shown Figure 2[12].

B. Pre-Processing

The obtained image cannot be used directly as it presents a set of weak features. Also the region of interest which is the nucleolus needs to be separated from the rest of the image. There are numerous methods proposed for doing this [1]-[11].

The pre-processing technique employed in this paper converts the RGB image into Gray scale colour model and then filters the noise with the help of a median filter and process the image using a set of morphological operators to strengthen the image. At the end we end up getting an image as shown in figure 3 [12][13].

C. Feature Extraction

There are several textural features available but in this paper the gray level co-occurrence matrix (GLCM) features, Haralick, gradient and tamura based features are taken for analysis.

GLCM is the simplest way of describing the features in a given image. It makes use of statistical moments present in the intensity moments of the intensity histogram of an image.

\[ P(i, j) = \sum_{x=1}^{N} \sum_{y=1}^{N} I(x, y) = i \land I(x+\Delta_x, y+\Delta_y) = j \]

(1)

Where \( i \) and \( j \) are the image intensity values, \( p \) and \( q \) are the spatial positions in an image \( I \) and the offset \( (\Delta_x, \Delta_y) \) specifies the distance between the pixel-of-interest and its neighbour.

In 1973 Haralick introduced 14 statistical features which are obtained by calculating the values for each one of the co-occurrence matrix obtained by averaging the four values using the directions 0°, 45°, 90° and 135°. The Symbol \( \Delta \) representing the distance parameter, can be selected as one or higher. In general \( \Delta \) can be set to 1 as the distance parameter. Where \( E \subset \mathbb{Z}^2 \times \mathbb{R} \) is the extinction image after intensity conversion.

\[ h = \text{histogram}(X(O)) \]

(2)

\[ h(v) = \text{frequency of pixel value } v \]

(3)

In Tamura features, in an Image we calculate a value for the three features at each pixel and treat these as a spatial joint coarseness-contrast-directionality (CND) distribution.

In this paper we have employed around 34 features for our analysis. The features implemented are listed in table I [14]-[18].

<table>
<thead>
<tr>
<th>Name of the feature</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean ( M1 )</td>
<td>( M1 = \frac{1}{M0} \sum_{v} v^4 h(v) = \sum_{v} v^4 H(v) )</td>
</tr>
<tr>
<td>Standard deviation ( M2 )</td>
<td>( M2 = \frac{1}{M0} \sum_{v} v^2 h(v) - (M1)^2 )</td>
</tr>
<tr>
<td>Median ( MED = \text{median}(X(O)) )</td>
<td></td>
</tr>
<tr>
<td>Entropy ( H(v) \text{log}_e(H(v)) \text{ entropy of } h )</td>
<td></td>
</tr>
<tr>
<td>Nucleocytoplasmic Ratio (NCR) ( NCR=Na/Ca )</td>
<td></td>
</tr>
<tr>
<td>Sum Variance ( f_i = \sum_{i=2}^{2N} (i - \bar{x})^2 P_{xy}(i) )</td>
<td></td>
</tr>
<tr>
<td>Contrast ( \sum_{i=0}^{N} P_{i,j} (i - j)^2 )</td>
<td></td>
</tr>
<tr>
<td>Hyperchromasia ( T_w(f) = f - f \circ b )</td>
<td></td>
</tr>
<tr>
<td>Sum ( M0 = {0} )</td>
<td></td>
</tr>
<tr>
<td>Skewness ( M3 = \frac{1}{(M2)^{3}} { \sum_{v} v^3 h(v) - 3 \sum_{v} v h(v) \sum_{v} v^2 h(v) + 2 \sum_{v} v^4 h(v) } )</td>
<td></td>
</tr>
<tr>
<td>Kurtosis ( M4 = \frac{1}{(M2)^{4}} { \sum_{v} v^4 h(v) - 3 \sum_{v} v^2 h(v) \sum_{v} v h(v) + 2 \sum_{v} v^3 h(v) } )</td>
<td></td>
</tr>
<tr>
<td>Max ( \max(X(O)) )</td>
<td></td>
</tr>
</tbody>
</table>

class of largest non – zero frequency
International Journal of Emerging Technology and Advanced Engineering  

<table>
<thead>
<tr>
<th>of h</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Autocorrelation</td>
<td>( f_6 = \sum_{i} \sum_{j} (i, j) (p(i, j)) )</td>
</tr>
<tr>
<td>Correlation</td>
<td>( \sum_{i=0}^{N-1} P_{ij} \left( \frac{(i-\mu_x)(j-\mu_y)}{\sqrt{(\sigma_x^2)(\sigma_y^2)}} \right) )</td>
</tr>
<tr>
<td>Cluster Prominence</td>
<td>( f_8 = \sum_{i} \sum_{j} (i + j - \mu_x - \mu_y)^2 p(i, j) )</td>
</tr>
<tr>
<td>Cluster shade</td>
<td>( f_7 = \sum_{i} \sum_{j} (i + j - \mu_x - \mu_y)^2 p(i, j) )</td>
</tr>
<tr>
<td>Dissimilarity</td>
<td>( f_6 = \sum_{i} \sum_{j}</td>
</tr>
<tr>
<td>Energy</td>
<td>( f_1 = \sum_{i} \sum_{j} [p(i, j)]^2 )</td>
</tr>
<tr>
<td>Entropy of GLCM</td>
<td>( f_9 = \sum_{i} \sum_{j} p(i, j) \log (p(i, j)) )</td>
</tr>
<tr>
<td>Homogeneity</td>
<td>( f_5 = \sum_{i} \sum_{j} \frac{1}{1+(i-j)^2} p(i, j) )</td>
</tr>
<tr>
<td>Maximum Probability</td>
<td>( f_{10} = MAX \sum_{i, j} p(i, j) )</td>
</tr>
<tr>
<td>Sum of average</td>
<td>( f_6 = \sum_{i=2}^{2N} i P_{xy}(i) )</td>
</tr>
<tr>
<td>Sum of square variance</td>
<td>( \sigma_x^2 = \sum_{i=0}^{N-1} P_{ij} (i-\mu_x)^2 ) ( \sigma_y^2 = \sum_{i=0}^{N-1} P_{ij} (j-\mu_y)^2 )</td>
</tr>
<tr>
<td>Sum of entropy</td>
<td>( f_8 = \sum_{i=2}^{2N} P_{x+y}(i) \log \left( P_{x+y}(i) \right) )</td>
</tr>
<tr>
<td>Difference variance</td>
<td>( f_{10} = \text{variance of } P_{x-y} )</td>
</tr>
<tr>
<td>Difference Entropy</td>
<td>( f_{11} = -\sum_{i=0}^{N-1} P_{x-y} (i) \log \left( P_{x-y}(i) \right) )</td>
</tr>
</tbody>
</table>

\[ f_{12} = \frac{HXY - HXY1}{\text{MAX} \{ HX, HY \}} \]

\[ f_{13} = (1 - \exp[-2.0(HXY2 - HXY)])^{1/2} \]

\[ HXY = -\sum_{i} \sum_{j} p(i, j) \log (p(i, j)) \]

\[ HXY1 = -\sum_{i} \sum_{j} p(i, j) \log (p_x(i)p_y(j)) \]

\[ HXY2 = -\sum_{i} \sum_{j} p_x(i)p_y(j) \log (p_x(i)p_y(j)) \]

\[ \sum_{i=1}^{G} \frac{C_{ij}}{1+i-j^2/G^2} \]

\[ \sum_{i=1}^{G} \frac{C_{ij}}{1+i-j^2/G^2} \]

\[ \sum_{i=1}^{G} \frac{C_{ij}}{1+i-j^2/G^2} \]

**Information Measurement of Correlation 1 & 2**

**Inverse difference**

**Inverse difference normalized**

**Inverse difference moment normalized**

**Coarseness**

<table>
<thead>
<tr>
<th>Coarseness</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1:-</td>
<td>For every point ( n_0, n_1 ) calculate differences between the not overlapping neighborhoods on opposite sides of the point in horizontal and vertical direction: and ( S^{k_1}(n_0, n_1) = \left( A_{n_0 + 2^k_1, n_1} - A_{n_0, n_1 - 2^k_1, n_1} \right) ) ( \Delta^k_1(n_0, n_1) = \left( A_{n_0 + 2^k_1, n_1 + 2^k_1} - A_{n_0 - 2^k_1, n_1 + 2^k_1} \right) )</td>
</tr>
<tr>
<td>Step 2:-</td>
<td>At each point ( n_0, n_1 ) select the size leading to the highest difference value: ( S(n_0, n_1) = \max_{k=1,...,5} \max_{d \in k_1} \Delta^k_d(n_0, n_1) )</td>
</tr>
<tr>
<td>Step 3:-</td>
<td>Finally take the average over 2S as a coarseness measure for the image: ( F_{crs} = \frac{1}{N_0N_1} \sum_{n_0=1}^{N_0} \sum_{n_1=1}^{N_1} 2^S(n_0, n_1) )</td>
</tr>
</tbody>
</table>

**Directionality**

\[ \theta = \frac{\pi}{2} + \tan^{-1} \frac{\Delta v(n_0, n_1)}{\Delta v(n_0, n_1)} \]

**Tamura feature**

**Contrast**

\[ F_{co} = \frac{\sigma}{\alpha^2} \text{ with } \alpha^2 = \frac{H_4}{\sigma^4} \]

91
Where

- \( Na \) is the number of pixels in the cell nucleolus region
- \( Ca \) is the number of pixels in the cytoplasmic region
- \( f \) is the grayscale image
- \( b \) is the structuring element
- \( o \) is the opening operation
- \( O \) is the total number of elements in the ROI
- \( h(v) \) frequency of pixel value \( v \)
- \( M0 \) is the total number of elements in the ROI
- \( M1 \) is the mean
- \( M2 \) is the standard deviation
- \( X(O) \) is the set of all pixels of \( X \) in \( O \) which is the class of largest non – zero frequency of \( h \)
- \( p(i,j) \) is the \( i,j^{th} \) entry in the normalized graytone spatial dependence matrix
- \( P(i,j)/R \) and \( N_{g} \) is the number of distinct gray levels in the quantized image.
- \( \mu_x, \mu_y, \sigma_x \) and \( \sigma_y \) are the means and standard deviations of \( p_x \) and \( p_y \)
- \( p_{x+y} = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} p(i,j) \) and \( N_g \) is the number of distinct gray levels in the quantized image.
- \( \text{Sum}_\text{Ent} \) is the Sum of Entropy features

- \( HX \) and \( HY \) are entropies of \( p_x \) and \( p_y \)
- \( \mu_x = \sum_i \sum_j p(i,j) \) \( \mu_y = \sum_i \sum_j p(i,j) \) \( p(i,j) \) be the \( (i,j)^{th} \) entry in normalized GLCM.
- \( C_{ij} = p_{ij} / \sum_i \sum_j p_{ij} \) in which \( p_{ij} \) represents the number of occurrences of gray levels \( I \) and \( j \) within the given window, given a certain \((\delta, \theta)\) pair \((\delta\text{-interpixel distance } \theta \text{-orientation}) \) and \( G \) is the quantized number of gray levels.
- \( \mu_z = \frac{\sum_{n_0=1}^{N_0} \sum_{n_1=1}^{N_1} X(n_0, n_1) - \mu}{4} \) is the fourth moment about the mean \( \mu \)
- \( \sigma^2 \) is the variance of the gray values present in the image and \( z \) has to be experimentally determined to be \( \frac{1}{4} \).

D. Herlev Dataset

In this paper we need to find a set of features that would work as a group to help find the better balance of features. The data set being used to test the hypothesis is the Herlev data set, which consists of 917 single pap smear images[19]. Cyto technicians and doctors have manually classified each cell into one of the 7 classes, namely superficial squamous, intermediate squamous, columnar, mild dysplasia, moderate dysplasia, severe dysplasia and carcinoma in situ. The first three correspond to normal classes and the last four classes correspond to abnormal classes as show in figure 4.

The pre-classification helps to serve as the gold standard for the entire work.

**Sample Images of Herlev Dataset**

**Normal Cells**

Superficial Squamous

Intermediate Squamous

Columnar

**Abnormal Cells**

Mild Dysplasia

Moderate Dysplasia

Severe Dysplasia

Carcinoma in Situ

Figure 4. Sample images of the Herlev Dataset

E. Outlier Deletion

The features mentioned in Table I are extracted from all the 917 images from the dataset and stored in the database. For each image there are about 34 features which have been extracted. Out of the 917 images there are about 7 groups of images or stages. So each 34 features for each image of the group are stored collectively. Out of the 34 features some are prominent and some are not. To find this we plot the values obtained in a graph. For eg:- take Auto correlation features and assume that all it has been calculated for all images in the superficial squamous stage. Now the values that represent close proximity will tend to cluster together and the ones that do not find close proximity will tend to move out of the clusters. These are called outliers and won’t be useful for detecting the particular stage of cancer. To help perform this grubs test is being used.
The test is based on the difference of mean and the most extreme data considering the standard deviation given dataset with assumed normal distribution. The test is based on the following formulas [20]

\[ T_{\text{max}} = \frac{x_n - X_{\text{mean}}}{S} \]  
\[ T_{\text{min}} = \frac{X_{\text{mean}} - x_1}{S} \]

Where
\[ x_n \text{ or } x_1 \text{= the suspected single outlier (max or min)} \]
\[ S \text{= standard deviation of the whole data set} \]
\[ X_{\text{average}} = \text{mean} \]

Once the outliers are deleted the total percentage of images present are calculated using the following formula

\[ \% \text{ of correct detection} = \frac{\Delta I}{\Delta S} \]

Where
\[ \Delta I \text{ is the total images present in a stage after deletion of outliers} \]
\[ \Delta S \text{ is the total images present in a the given dataset for particular stage} \]

Using the above formulas we plot each and every feature value for each image of a particular feature of a particular stage. In Figure 5 a sample for the Nucleo cytoplasmic ratio feature for Superficial Squamous stage is shown. As you can see the the ones that do not work well for this feature are far away from the cluster and using this we can calculate the percentage of success for each feature for each stage.

F. Ranking

Once the features have been extracted and the outliers are deleted we are left with a percentage count. Now we can just do a normal rank and take the top 5 or you can take values above 90%. But doing so sacrifices the efficiency of the overall system. Hence a unique ranking system is proposed. The features in the database are listed after outlier detection as columns representing stages and rows representing features. Now as we mentioned above there may be many probable’s and to scrutinize this we use a Gaussian function. Using this function a bell curve is drawn over the values after segregating them into bins as shown in Figure 6. It is calculated using the following equation

\[ f(x) = a \exp \left( -\frac{(x-b)^2}{2\sigma^2} \right) \]  

III. EXPERIMENTAL RESULTS

The above discussed methodologies where successfully executed on all the 917 images of the Herlev dataset in the quest to attain a better ranking technique to produce a combinational set of features. A sample of images segmented using the segmented techniques discussed above is shown in figure 7.
The features discussed in Table I are extracted and the outliers are determined and deleted and the percentage of success is calculated and is shown in table II. Then the values left after outlier deletion are taken and are segregated into bins and a bell curve is drawn over them. This is shown in figure 8. For each images the bell curve is drawn. These bins are so arranged that the ones with less proximity are clearly visible. By setting a maximum and minimum threshold values we can obtain the necessary values. Here in this work we are not mentioning the threshold value which has to be weighed between accuracy and speed. Unlike segmentation algorithms where threshold can be defined here it depends on factors like processing speed, need of the client. From the graph it is clear that the bins in the middle will be the high proximity values that can create a combinational set of features.

Figure 7. Carcinoma-in-situ image segmented

The percentage of values after removing outliers is shown in table II

<table>
<thead>
<tr>
<th>TABLE II</th>
<th>Percentage of liable images after outlier deletion</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>S1</td>
</tr>
<tr>
<td>1</td>
<td>90.54</td>
</tr>
<tr>
<td>2</td>
<td>82.43</td>
</tr>
<tr>
<td>3</td>
<td>85.14</td>
</tr>
<tr>
<td>4</td>
<td>74.32</td>
</tr>
<tr>
<td>5</td>
<td>81.08</td>
</tr>
<tr>
<td>6</td>
<td>81.08</td>
</tr>
<tr>
<td>7</td>
<td>79.73</td>
</tr>
<tr>
<td>8</td>
<td>83.78</td>
</tr>
<tr>
<td>9</td>
<td>81.08</td>
</tr>
<tr>
<td>10</td>
<td>87.84</td>
</tr>
<tr>
<td>11</td>
<td>72.97</td>
</tr>
<tr>
<td>12</td>
<td>85.14</td>
</tr>
<tr>
<td>13</td>
<td>89.19</td>
</tr>
<tr>
<td>14</td>
<td>90.54</td>
</tr>
<tr>
<td>15</td>
<td>87.84</td>
</tr>
</tbody>
</table>

Where

- Abnormal Cells
  - S1: Carcinoma in Situ
  - S2: Mild Dysplasia
  - S3: Moderate Dysplasia
  - S4: Severe Dysplasia

- Normal Cells
  - S5: Columnar Dysplasia
  - S6: Intermediate Squamous
  - S7: Superficial Squamous

Note: The order of the stages has been changed to help in computation.

Features 1 to 34 represent the same order as in Table I
Figure 8 Gaussian Bell Curve for the Horizontal Ranking
(a) Normal Superficial (b) Normal Intermediate (c) Normal Columnar (d) Light Dysplasia (e) Moderate Dysplasia (f) Severe Dysplasia (g) Carcinoma in Situ

Note: F1 to F34 are in the same order as the features ordered in Table I.

IV. CONCLUSION

The proposed methodologies are able to create a combinational set of features by using Gaussian bells curve drawn using a Gaussian function. Using this we can balance the speed and accuracy quotient of the feature set. In future as processing speed increase other features can also be analysed. Features like wavelets, MPEG-7 feature descriptors can be used to better the odds of detecting the cancer with balanced speed and accuracy.

REFERENCES


[16] Prof. Dr.-Ing.H.Ney., Features for Image retrieval, December 2003


