Wavelet Based Texture Features of Cancer Cytology Images

Mrs. Ancy David S

Dept. of Computer Science and Engineering College of Engineering, Karunagapally, Kerala, India

Abstract - The early detection of lung cancer is a challenging problem, due to the structure of the cancer cells, where most of the cells are overlapped with each other. This paper presents a novel method for extracting features of the cell from sputum colour images to detect the lung cancer in its early stages. The manual analysis of the sputum samples is time consuming, inaccurate and requires intensive trained person to avoid diagnostic errors. The results will be used as a base for a Computer Aided Diagnosis (CAD) system for early detection of lung cancer which will improve the chances of survival for the patient. Sputum cytology is a non-invasive method used in computer aided earlier lung cancer detection because it is cost effective and not painful compared to biopsy method. Also other imaging modalities like CT-images cause radiation effect and also getting CT-images is time consuming. In the proposed method first to eliminate unwanted information from the stained image and then level set method is used for segmenting the cells from cytoplasm and then to extract features from the cell a comparative study is made to find the best feature extraction method. The methods used are Grey Level Co-occurrence Matrix (GLCM), Discrete Wavelet Transform (DWT) and Stationary Wavelet Transform (SWT). And then classification is done using Support Vector Machine (SVM) classifier. From the experimental results it is found that by using SWT the accuracy of the system increased to 94%.

Key Words: Sputum cytology, Discrete Wavelet Transform, Stationary Wavelet transform, Support Vector Machine.

1. INTRODUCTION

Cancer is a class of diseases characterized by out-of-control cell growth. There are over 100 different types of cancer, including breast cancer, skin cancer, lung cancer, colon cancer, prostate cancer, and lymphoma, and each is classified by the type of cell that is initially affected. Compared with other types of cancer, the outlook for lung cancer is not very good. Overall, of all types of lung cancer, about 32 out of every 100 people (32%) will live for at least 1 year after they are diagnosed. Around 10 out of every 100 people (10%) will live for at least 5 years. And about 5 out of every 100 people (5%) will live for at least 10 years. Lung cancer has been the largest cause of cancer deaths worldwide, since its symptoms can be found exclusively in advanced stages where the chances for patients to survive are very low, thus making the mortality rate the highest among all other types of cancer.

In this work sputum cytology images are used, since it is known as the golden standard method. Several previous studies suggest that sputum cytology is the best and affordable method for lung cancer detection [1][2]. Probably the cancer might be detected when the cell has already been metastasized; in that stage it is difficult to get treated. Early diagnosis of lung cancer dramatically improves survival rates for this disease. In this paper, Pap smear sputum cytology images in RGB color space are acquired preprocessed to enhance quality of the image and to remove background. Then the nuclei are extracted from the image and the dominant features are extracted from these nuclei. Features extracted from these nuclei serve to diagnosis of cancer that means the benign and malignant cells can be classified. And in the final phase using a classifier the benign and malignant cells are classified. Later the efficiency of this system is calculated.

The major contribution of this paper is on feature extraction phase. Texture Analysis is important in image processing for the segmentation and classification of images based on local spatial variations of intensity. In this paper we are concerned with the target of developing a set of textural features using several feature extraction methods. The wavelet transform has been shown to be a useful tool for analyzing textures and it plays important role in biomedical applications. Discrete wavelet transforms (DWT) based texture features are used to classify cancerous cells. GLCM is the statistical feature extraction method, which also gives textural information. We can’t say it is efficient method since features are not invariant to rotation or scale changes in the texture. Another one is SWT. The main drawback of DWT is that it is shift invariant. SWT can overcome this limitation and it provides comparatively higher accuracy than other two methods. Inspired by this information the objective of this work is to design an automated lung cancer detection system for the early detection of lung cancer using image processing techniques.

2. LITERATURE SURVEY

Diagnosis of cancer in image processing involves mainly four steps: (a) Preprocessing (b) Segmentation (c) Feature Extraction and (d) Classification. Many research papers with different approaches for diagnosis of lung cancer in its early stages reported in the literature. The key objective is to highlight key strengths and limitations to these techniques.

ISO 9001:2008 Certified Journal | Page 2991
Y.M.George et al.[3] proposed Histogram Stretching followed by Contrast Limited Adaptive Histogram Equalization due to the low quality of Cytological images which helps to improve the quality of image. M.Hrebien et al.[4] proposed simple histogram with a linear transformation of the image levels of intensities in the FNAC images as an enhanced technique to improve the quality of the image. This contrast correction in each color channel gives better result for the subsequent stages. Many researches has been done to extract either cell region or nucleus region from the sputum images. F.Taher et al.[5] presented a framework for segmentation and extraction of sputum cells in sputum images using Bayesian classification method followed by region detection method. And in another work of same authors for the cell detection the comparison between threshold technique and Bayesian classification is made and found out that Bayesian method achieved the best scores and later for cell extraction mean shift segmentation algorithm is used. Followed by this work same authors[6] presented two methods for detecting sputum cells they are, rule based method and Bayesian classification. And they suggested that Bayesian classification to be an appropriate for the segmentation. J.Malek et al.[7] proposed a segmentation method called GVF-Snake algorithm for the isolation of the cells in breast cancer cytological image. This method outperforms ordinary Snake algorithm. L.Jelen el at.[8] made use of automated segmentation procedure called level set method for segmentation. It shows best results than Hough Transform. To analyse the given cell is cancerous or not is based on features extraction from the segmented cell. The important one is nuclear features. The nuclear feature such as Area, Perimeter, Convexity and Eccentricity are extracted in [8]. These defines the shape and size of the nuclei. A.K.Mohanty et al.[9] they use GLCM and GLRLM feature extraction method. After comparison either it is ultrasound image or cytology image GLCM is found to be more accurate. A comparison between Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) is made in [10] and DWT is found to be more accurate in classification stage. The DWT features such as mean, second central moment, Entropy are the features extracted from segmented nuclei in [7]. The Complex wavelets based features were also found best accuracy in diagnosis of cancer [11][12][13]. I.Niwas et al.[11] proposed basic algorithm for evaluating the variability of clustered nucleus in breast cytology image using Complex Bi-orthogonal Wavelet bases without segmentation and yield good accuracy but this method is limited to number of samples. Again they applied Dual-tree complex wavelet transform [12] on color images in another work. By using the application of Complex Daubeics Wavelet[13] on unsegmented nuclei with limited number of samples yield good result. Finally in [14] they extended their work to segmented nuclei and yield better result. The final stage is to make a decision whether the cell is cancerous or not.

This is done by using a Classifier. Based on the complexity of the data the classifier is designed. K-nearest neighbour classifier is used in [15][14]. It is the most popular technique for classification. J.Malek et al.[7] used Fuzzy c-means clustering for the classification and good rate of classification is obtained. In [16] they evaluated the performance of fuzzy c-means and SVM, and the selection of any one of these algorithm is based on the purpose of detection of nuclei in the image. ie, if the purpose is to find malignant FCM may be preferable and to find true nuclei SVM is preferred. SVM is also used in [17] for the classification and it is compare with Naive Bayes and Artificial Neural Network ad found out that accuracy of the system increased with SVM classifier.

3. MATERIALS AND METHODS

A. Materials

The data we are using here is sputum images which are collected by means of Sputum cytology, it examines a sample of sputum (mucus) under a microscope to determine whether abnormal cells are present. In this work the Database of 2D sputum color images is being used, which contains normal and abnormal cases. These sputum samples have been stained, by using Papanicolaou standard staining methods. Pap staining is a very reliable technique. The entire procedure is known as Pap smear.

B. Methods

Image processing tools can be used in medical application. One such application is cancer diagnosis. It involves four steps:

(i) Preprocessing

The pre-processing of image is for selectively removing the noise present in the images without affecting the details. It plays key role in the diagnostic and analysis process. Preprocessing is the important method that influences automated detection of defects. There are several preprocessing algorithms are there to improve the quality of the image and make the subsequent phases as an easier and reliable one. In cytology images, enhancement is required. So i went for image enhancement algorithm known as Recursive Mean Separate Histogram Equalization(RMSHE), which helps to improve contrast of the image. It is an extension of Histogram Equalization method. Instead of decomposing an image only once, this method proposes decomposing image recursively up to a scalar r. Therefore 2 sub images will be generated. Then each sub image equalized separately using Histogram Equalization method. If r=0, that means no sub image decomposition is done, which is equivalent to HE method. When one mean separation is done before equalization, i.e. r=1, this is equivalent to BBHE. This increases a level of brightness preservation. Similarly, two mean-separations before equalization will
result in much higher level of brightness preservation as compared to r=0 and r=1 levels. Thus we can conclude that level of brightness preservation will increase with the increase of number of recursive mean separations. This technique aims to bring more extends of brightness preservation than HE and BBHE techniques.

(ii) Segmentation

Segmentation of an image is to divide an image into several images. Segmentation can be done on the basis of the application. In our medical images, more precisely in lung cytology image, segmentation is the process of extracting cell region from the sputum image. Likewise, based on Region of Interest (ROI) the segmentation can be performed. The goal of segmentation is to change the representation of image into something meaningful one for easier analysis.

Level Sets are an important category of modern image segmentation techniques are based on partial differential equations (PDE), i.e. progressive evaluation of the differences among neighboring pixels to find object boundaries. Ideally, the algorithm will converge at the boundary of the object where the differences are the highest.

Level set methods using level sets as a tool for numerical analysis of surfaces and shapes. The advantage of the level set model is that one can perform numerical computations involving curves and surfaces on a fixed Cartesian grid without having to parameterize these objects. In this work we make use of the automated segmentation procedure that involves the level set method. The level set approach allows the evolving front to change topology, break, and merge, which means that the evolving front can extract the boundaries of particularly intricate contours. In addition, the method works in three dimensions with almost no change, so three dimensional surfaces can be extracted as well.

(iii) Feature Extraction

Feature is the method of extracting the most prominent feature in an image for analysis. Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. When performing analysis of complex data one of the major problems stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power or a classification algorithm which overfits the training sample and generalizes poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy. The extracted features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data. Color, Texture, Shape are the important features in an image.

In this work, the importance giving to texture features. Texture tactile or visual characteristic of a surface. Texture analysis aims in finding a unique way of representing the underlying characteristics of textures and represent them in some simpler but unique form, so that they can be used for robust, accurate classification and segmentation of objects.

Discrete Wavelet Transform (DWT) is the one of the spatial feature extraction method that we are using here.

Discrete Wavelet Transform (DWT) is any wavelet transform for which the wavelets are discretely sampled. As with other wavelet transforms, a key advantage it has over Fourier transforms is temporal resolution: it captures both frequency and location information (location in time). The discrete wavelet transform (DWT) is a linear transformation that operates on a data vector whose length is an integer power of two, transforming it into a numerically different vector of the same length. The DWT of a signal x is calculated by passing it through a series of filters. First the samples are passed through a low pass filter with impulse response g resulting in a convolution of the two:

\[ y[n] = (x * g)[n] = \sum_{k=-\infty}^{\infty} x[k]g[n-k] \]  

(1)

The signal is also decomposed simultaneously using a high-pass filter h. The outputs giving the detail coefficients (from the high-pass filter) and approximation coefficients (from the low-pass) is inherently important that the two filters are related to each other and they are known as a quadrature mirror filter. Wavelets seem to be effective for analysis of textures recorded with different resolution.

Stationary Wavelet Transform

We know that the classical DWT suffers a drawback: the DWT is not a time-invariant transform. This means that, even with periodic signal extension, the DWT of a translated version of a signal X is not, in general, the translated version of the DWT of X.

The Stationary wavelet transform (SWT) is a wavelet transform algorithm designed to overcome the lack of translation-invariance of the discrete wavelet transform (DWT). Translation-invariance is achieved by removing the down samplers and up samplers in the DWT and up sampling the filter coefficients by a factor of \(2^j\) in the \(j^{th}\) level of the algorithm. The SWT is an inherently redundant scheme as the output of each level of SWT contains the same number of samples as the input – so for a
decomposition of N levels there is a redundancy of N in the wavelet coefficients.

**Implementation**

The following block diagram depicts the digital implementation of SWT.

![Figure 1 SWT Decomposition](image)

**Figure 1 SWT Decomposition**

In this study, we investigated eight textural features to quantitatively evaluate the textural characteristics of the lung cells on cytology images. These features were selected based on experimental observations. The eight textural features are defined as follows.

1. **Angular second moment (ASM, Energy)**
   
   \[
   \text{ASM} = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} p(i,j)^2
   \]
   
   ASM, also known as energy, is a measure of homogeneity of an image. A homogeneous scene will contain only a few gray levels, so that GLCM will have a few but relatively high values of \(p(i,j)\). Thus, the sum of squares will be high. It is the monotonic gray-level transition; higher values indicate textural uniformity. Therefore, when the image is homogeneous, the ASM will have high values.

2. **Inverse difference moment (IDM)**
   
   \[
   \text{IDM} = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \frac{1}{(i-j)^2} p(i,j)
   \]
   
   IDM measures the local homogeneity of an image. The incidence of co-occurrence of pixel pairs is enhanced when they are close in gray-scale value and thus increases the IDM value. Because of the weighting factor \((1+(i-j)^2)^{-1}\), it will get small contributions from inhomogeneous areas \(i \neq j\). The result is a low IDM value for inhomogeneous images, and a relatively higher value for homogeneous images.

3. **Contrast**
   
   \[
   \text{CONTRAST} = \sum_{n=0}^{G-1} n^2 \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} p(i,j), \quad |i-j| = n
   \]
   
   Contrast is a measure of the local variations presented in an image. This measure of contrast favors contributions from \(p(i,j)\) away from the diagonal. It is highly correlated with the difference between the highest and the lowest values of a continuous set of pixels, particularly when the value of the displacement vector is 1. If there is a large amount of variation in an image, the contrast will be high.

4. **Variance**
   
   \[
   \text{VARIANCE} = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} (i-\mu_x)^2 p(i,j)
   \]
   
   Variance puts relatively high weights on the elements that differ from the average value of \(p(i,j)\). It refers to the gray-level variability of the pixel pairs and is a measurement of heterogeneity. Variance increases when the gray-scale values differ from their means. Unlike contrast, variance has no spatial frequency. Although a high variance is suggestive of a high contrast value, the converse relationship does not apply.

5. **Correlation**
   
   \[
   \text{CORRELATION} = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \frac{(i \times j) \times p(i,j) - \mu_x \times \mu_y}{\sigma_x \times \sigma_y}
   \]
   
   Correlation is a measure of the linear dependency of gray levels on those of neighboring pixels or specified points. It indicates local gray-level dependency on the texture image; higher values can be obtained for similar gray-level regions.

6. **Entropy**
   
   \[
   \text{ENTROPY} = -\sum_{i=0}^{G-1} \sum_{j=0}^{G-1} p(i,j) \times \log(p(i,j))
   \]
   
   Entropy measures the randomness of the image texture (intensity distribution). Entropy is the highest when all the probabilities \(p(i,j)\) are equal, and smaller when the entries in \(p(i,j)\) are unequal. Therefore, a homogeneous image will result in a lower entropy value, while an inhomogeneous (heterogeneous) region will result in a higher entropy value.

7. **Cluster shade**
   
   \[
   \text{SHADE} = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} (i+j-\mu_x-\mu_y)^3 \times p(i,j)
   \]
Cluster shade is a measure of the skewness of the matrix and is believed to gauge the perceptual concepts of uniformity. A new “i+j” image is created, having a range of integer intensities from 0 to 2(N\_g - 1). The u\_i\_j value is computed and stored for the first neighborhood of the image, and is subsequently updated as the neighborhood is moved by one pixel. When the cluster shade is high, the image is asymmetric.

8. Cluster prominence

\[ PROM = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} (i + j - \mu_x - \mu_y)^4 \times P(i,j) \]  

Cluster prominence is also a measure of asymmetry. When the cluster prominence value is high, the image is less symmetric. In addition, when cluster prominence value is low, there is a peak in the GLCM matrix around the mean values. For a cytology image, a low cluster prominence value indicates small variation in gray-scale.

9. Kurtosis

\[ KRUTOSIS = \frac{\sum_{i=1}^{N} (\mu_i - \mu)^4}{s^4} \]  

The height and sharpness of the peak relative to the rest of the data are measured by a number called kurtosis. Higher values indicate a higher, sharper peak; lower values indicate a lower, less distinct peak. Thus in cytology images the more intensity values can be calculated, which can be chromatin feature.

10. Dissimilarity

\[ DISS = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} |P(i-j)| \]  

In the Contrast measure, weights increase exponentially (0, 1, 4, 9, etc.) as one moves away from the diagonal. However in the dissimilarity measure weights increase linearly (0, 1, 2, 3 etc.).

11. Homogeneity

Dissimilarity and Contrast result in larger numbers for more contrast windows. If weights decrease away from the diagonal, the result will be larger for windows with little contrast. Homogeneity weights values by the inverse of the Contrast weight, with weights decreasing exponentially away from the diagonal.

12. Mean

\[ MEAN = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} iP(i,j) \]  

The pixel value is weighted not by its frequency of occurrence by itself (as in a “regular” or familiar mean equation) but by its frequency of its occurrence in combination with a certain neighbour pixel value.

13. Standard Deviation

\[ SD = \sqrt{\sum_{i=0}^{G-1} \sum_{j=0}^{G-1} (P(i,j) - \mu)^2} \]  

14. Autocorrelation

An autocorrelation function measures the linear spatial relationships between spatial sizes of texture primitives. Autocorrelation-based approach to texture analysis is based on the intensity value concentrations on all or part of an image represented as a feature vector. Calculation of the autocorrelation matrix involves individual pixels.

15. Sum of Squared Difference and Sum Of Absolute Differences

Sum of Squared Difference (SSD) and Sum of Absolute Difference (SAD) are used to measure the similarity between train image and the test images.

Likewise nineteen textural features were extracted using these three methods like energy, entropy, contrast, local homogeneity, correlation, and shade, and provenance, sum of squares, sum average, sum entropy, difference entropy, sum variance and difference variance.

(iv) Classification

Classification is the process of classifying the given image is cancerous or not. The selection of classifier for medical diagnosis application depends on the sample size and selection of relevant features from the data set. Classification is a task of assigning an item to a certain category, called a class, based on the characteristic features of that item. This task in any classification system is performed by a classifier that takes a feature vector as an input and responds with a category to which the object belongs. A feature vector is a set of features extracted from the input data.

In this work Support Vector Machine (SVM) classifier is used for classification. A (SVM) performs classification by constructing an N-dimensional hyperplane that optimally
separates the data into two categories. The goal of SVM modelling is to find the optimal hyperplane that separates clusters of vector in such a way that cases with one category of the target variable are on one side of the plane and cases with the other category are on the other side of the plane. The vectors near the hyperplane are called the support vectors. An SVM analysis finds the hyperplane that is oriented so that the margin between the support vectors is maximized.

4. EXPERIMENTAL RESULTS
The proposed method is tested over cytological images through MATLAB. The stained images first preprocessed using RMHSE method in order to eliminate unwanted information. Next step is to segment the cell region from this sputum image. For that purpose the most powerful medical image segmentation method called level set method is used. It finds the outer boundary of the cell region. So that we can segment cells more accurately. After segmentation, the main part of this work is feature extraction phase. Here I have extracted features from this segmented image and studied several statistical texture feature extraction techniques such as GLCM, DWT, SWT and comparisons were made between these three methods. In each of these methods 19 textures features are computed. After that using SVM classifier used for the classification.

![Fig. 2 (a)Sample sputum image (b)Preprosessed Image](Fig. 2)

![Fig. 3 Segmented Image](Fig. 3)

From the experimental results we can clearly see that SWT gives more accurate results than the other two methods. And the accuracy of the proposed system is increased when compared with existing systems.

5. CONCLUSION
In this work I have tried to develop an automated cell cancer detection system using image processing techniques. So for the early detection of lung cancer, first the image preprocessing will be done using RMHSE method, and then this enhanced image is used for cell segmentation. For segmentation the most efficient medical segmentation technique called level set method will be used. After extracting the nucleus the prominent features are extracted using GLCM, DWT and SWT. A comparison is made between these methods and it is found that SWT is the best feature extraction method for the early detection of lung cancer. By using this feature vector, through SVM classifier the decision can be made whether the cell is malignant or not. The performance results shows that this system dominates existing systems. The accuracy level can
be improved by using complex wavelets since it gives phase information. Here SVM classifier is used we can employ this system into new classifiers to improve the efficiency of the system.

REFERENCES


