International Journal of Advance Engineering and Research Development (IJAERD)
Volume 1,Issue 6,June 2014, e-ISSN: 2348 - 4470 , print-ISSN:2348-6406

Detection of Lung Cancer using Sputum Image Segmentation.

Dharmesh A Sarvaiya[1], Prof. Mehul Barot[2]

[1,2] *Department of Computer Engineering, L.D.R.P Institute of Technology & Research, KSV University, Gandhinagar*
dharmeshsarvaiya7@gmail.com

Abstract—Lung cancer is acknowledged to be the fundamental driver of disease passing worldwide, and it is difficult to detect in its early stages because symptoms appear only in the advanced stages causing the mortality rate to be the highest among all other types of cancer. The early detection of cancer can be helpful in curing disease completely. This research paper summarizes various reviews and technical articles on Lung cancer detection using the data mining techniques to enhance the Lung cancer diagnosis and prognosis and also deals with K Means image segmentation technique in MATLAB. With the help of different aspects of Nuclei of sputum images, such as size, shape, and area , Lung cancer can be detected by it. This report includes the survey of different techniques such as threshold classifier, a Bayesian classification , Fuzzy C Means and propose a method which uses the concept of k means algorithm. K Means algorithm is used to segment the sputum image into Nuclei, Cytoplasm and Debris Cells.

Keywords- Lung cancer detection, sputum images, threshold technique, Bayesian classification, Fuzzy C Means Algorithm, K Means algorithm.

I. INTRODUCTION

Lung Cancer, particularly the threatening sort is one of the deadliest malignancies. Throughout the last few years the occurrence of harmful tumor has constantly expanded, on the grounds that the cure of the disease depends very on its initial judgment emulated by a suitable surgical extraction. There are two major types of lung cancer, Non-small cell lung cancer (NSCLC) and Small cell lung cancer (SCLC)[1].

Staging lung cancer is based on whether the cancer is local or has spread from the lungs to the lymph nodes or other organs. Because the lungs are large, tumors can grow in them for a long time before they are found. Even when symptoms—such as coughing and fatigue—do occur, people think they are due to other causes. For this reason, early-stage lung cancer (stages I and II) is difficult to detect. Most people with lung cancer are diagnosed at stages III and IV[1].

Doctors utilize a few methods to diagnose lung tumor, for example, X-rays, CT Scan, PET scan etc. Furthermore, the statistics from the World Health Organization (WHO), indicate that deaths caused by cancer will reach about 12 million people in 2030[2].
The symptoms of lung cancer can include [3]:

- Coughing, especially if it persists or becomes intense
- Pain in the chest, shoulder, or back unrelated to pain from coughing
- A change in color or volume of sputum
- Shortness of breath
- Changes in the voice or being hoarse
- Harsh sounds with each breath
- Recurrent lung problems, such as bronchitis or pneumonia
- Coughing up phlegm or mucus, especially if it is tinged with blood

- Coughing up blood

A. Sputum Cytology[4]

Sputum cytology examines a sample of sputum (mucus) under a microscope to determine whether abnormal cells are present. Sputum is not the same as saliva. Sputum is produced in the lungs and in the airways leading to the lungs. Sputum has some normal lung cells in it.
Sputum cytology may be done to help detect certain non cancerous lung conditions. It may also be done when lung cancer is suspected. Sputum images are collected for the purpose of detecting disease in earlier stage.

A sputum sample may be collected [4]:

- By a person coughing up mucus.
- By breathing in a saltwater (saline) mist and then coughing.
- During bronchoscopy which uses a bronchoscope to look at the throat and airway.

The automatic of a sputum cell state is based on the analysis of its nucleus and cytoplasm. The sputum cells are characterized by uncertainty cells pattern that make the segmentation and detection of the cells very problematic, so it is difficult to segment the foregrounds from the image automatically and perfectly.

These sputum images are stained with two types [5].

- Type1, blue dye images resulting in the dark-blue nucleus of all the cells present in the image and clear-blue cytoplasm.

- Type 2, red dye images resulting in the dark-blue nucleus of the small debris cells with their corresponding small clear-blue cytoplasm regions, and red sputum cell with dark-red nucleus and clear-red cytoplasm.

II. RELATED WORK

For Early detection of Lung cancer , there are different approaches of data mining to detect nuclei of the sputum image and cluster those images.

In [5], researchers have proposed a Computer Aided Diagnosis (CAD) system for early detection of malignant lung cancer cells using digital images of stained sputum smears. Such an automated system would allow objective and unbiased assessment, as compared to human evaluation which might be corrupted by errors originating from inter-and intra-observer variability that characterizes human observation. Eventually, this system will be useful for handling large sputum image databases and relieving the pathologist from tedious and routine task.. In this paper, they focus on the extraction and segmentation of sputum cells from background regions. The sputum images are stained according to the Papanicolaou standard staining method .

In this paper they proposed two methods for addressing this problem the first employed a threshold-based technique. The second method uses a Bayesian classification framework. The problem of extracting the nucleus and the cytoplasm is approached using a combination of robust mean shift segmentation and rule-based techniques.

A. Sputum Cell Detection[5]

The unit discovery points at the extraction of the cell area from the sputum picture. This is carried out by verifying whether a pixel in the sputum picture fits in with the sputum cell utilizing its shade data. The staining strategy, connected in the sputum specimen result, permits, to some level, the sputum cell to have an unique chromatic manifestation opposite the foundation. In any case, the way of the sputum shade pictures, which holds numerous trash units and the relative complexity around the cytoplasm and cores cells, implies that the extraction process for the cores and cytoplasm units is not a direct system.

1). Threshold Technique [5].

The threshold technique depends on the staining methods by which the image is organized and derived from the difference in the brightness level in RGB components of the sputum color images.

Here The parameter Θ is determined by trial and error testing whereby the outcome of the segmentation is assessed visually.

For the image stained with blue dye, the following rule is used to extract sputum pixels:

If (B(x, y) < G(x, y) + Θ) then B(x, y) is a sputum
else B(x, y) is non sputum .

For the image stained with red dye, where the red colour is the most dominant colour between the sputum cells and the background, we use the following sequence of rules:

If (B(x, y) < G(x, y) or (B(x, y) > R(x, y)) then
B(x, y) is sputum else B(x, y) is non sputum.

The optimal value of the threshold was determined by analyzing the performance of the method across.

2). Bayesian Classification [5]

In this approach they address the cell detection problem using a probabilistic method based on the Bayesian classification.

In these methods, a pixel x is considered part of the sputum region if $p(bg|x) < p(sp|x)$ where sp and bg refer to the sputum and the background respectively. Applying the Bayesian Rule and the concept of classification cost, this inequality can be brought to [5]:

$$\sigma = \frac{\mu sp}{\mu bg}\frac{p(bg)}{p(sp)} < \frac{p(x|sp)}{p(x|bg)} \qquad (1.1)$$

where is μ_{sp} the loss weight incurred if the sputum class has been selected instead of the background and μ_{bg} is the loss weight incurred if the background class has been selected instead of the sputum, $p(bg)$ and $p(sp)$ are the probabilities of the background and the sputum classes respectively, and

they are estimated from the total number of sputum and background pixels in the training set of images according to the following equations:

$$p(\text{sp}) = \frac{Tsp}{Tsp+Tbg} \quad (1.2)$$

$$p(\text{bg}) = \frac{Tbg}{Tsp+Tbg} \quad (1.3)$$

where *Tsp* and *Tbg* are the numbers of sputum and background color respectively.

A database of images, they collected from the Tokyo Center for lung cancer, was utilized in this study. The size of each image is 768 × 512 pixels and they were provided in the RGB space. They conducted a comprehensive set of experiments to study the outcome of the threshold algorithm for the detection and extraction of the cells into sputum cells and background. Furthermore, they analyzed the essence of color representation and color quantization on the sputum cell detection. Then they used the cell extraction techniques the Bayesian classifier.
The Bayesian classification achieved the best scores. It succeeded particularly in reducing the number of False Negative and improving the sensitivity.

In [6], the researchers have proposed two segmentation methods, Hopfield Neural Network (HNN) and a Fuzzy C-Mean (FCM) clustering algorithm, for segmenting sputum color images to detect the lung cancer in its early stages. The segmentation 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. we applied a thresholding technique as a pre-processing step in all images to extract the nuclei and cytoplasm regions, because most of the quantitative procedures are based on the nuclear feature. The HNN and FCM methods are designed to classify the image of *N* pixels among *M* classes.

In thresholding technique, the filtering algorithm uses the appropriate range of the threshold parameter Θ which will allow an accurate extraction of the region of interest (ROI) composed of the nuclei and cytoplasm pixels. Thresold technique is detailed as earlier in this survey paper.The parameter Θ is determined by trial and error testing whereby the outcome of the segmentation is assessed visually.

3) . Fuzzy Clustering [5]

Fuzzy Clustering has been used in many fields like pattern recognition and Fuzzy identification. A variety of Fuzzy clustering methods have been proposed and most of them are based upon distance criteria. The most widely used algorithm is the Fuzzy C-Mean algorithm (FCM), it uses reciprocal distance to compute fuzzy weights. This algorithm has as input a pre-defined number of clusters, which is the *k* from its name. Means stands for an average location of all the members of particular cluster and the output is a partitioning of *k* cluster on a set of objects. The objective of the FCM cluster is to minimize the total weighted mean square error :

$$J = (W^{q^k}, C^{(k)}) = \sum_{(q=1,Q)} \sum_{(k=1,K)} (W_{qk})^p \, \| x^{(q)} - c^{(k)} \|^2$$

The FCM allows each feature vector to belong to multiple clusters with various fuzzy membership values. Then the final classification will be according to the maximum weight of the feature vector over all clusters.

Authors applied the FCM clustering algorithm with the specification mentioned sputum color images and maintain the results for further processing in further steps. FCM algorithm could segment the images into nuclei, cytoplasm regions and clear background, however, the FCM is not sensitive to intensity variation, therefore, the cytoplasm regions are detected as one cluster when they fixed the cluster number to three, four, five and six. Moreover, FCM failed in detecting the nuclei, it detected only part of it. By experiment,

III. PROPOSED TECHNIQUE.

For Early detection of Lung cancer using sputum cytology, the main focus of this research is on proper and efficient clustering techniques. With effective method of K Means for clustering sputum image result is more accurate in segmentation of sputum image. K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. K-means clustering algorithm is a simple clustering method with low computational complexity [7]. The clusters produced by K-means clustering do not overlap. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori.

The main idea is to define K centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other.

The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early grouping is done K-means clustering algorithm is an unsupervised method. It is used because it is simple and has relatively low computational complexity. In addition, it is suitable for biomedical image segmentation as the number of clusters (K) is usually known for images of particular regions of human anatomy[7].

- Advantages of K Means :
  - Cluster does not overlap.
  - Low Computational Complexity

- Disadvantage of K Means :
  - Requires specifying number of clusters to be partitioned.
  - Requires to specify number of colors to be present in sputum image.
  - Only segments colored image.

In the system for segmenting sputum images, stained sputum images are used as an input. These images are either stained with red dyes or with blue dyes. When Stained with Red dyes, the sputum cell nucleus becomes dark red with clear red cytoplasm region and debris cell nucleus becomes dark blue with clear blue cytoplasm. After that these stained sputum images are kept under microscope and they are magnified and Images are taken through Digital Camera fitted with Microscope. Now all these images will be used as input.

In MATLAB , with the use of inbuilt K Means function and color aspects, segmentation of Sputum image is done and Nuclei is extracted as a result to check whether it is cancerous or not based on shape and size of it.

Proposed Technique to segment the Sputum Image:

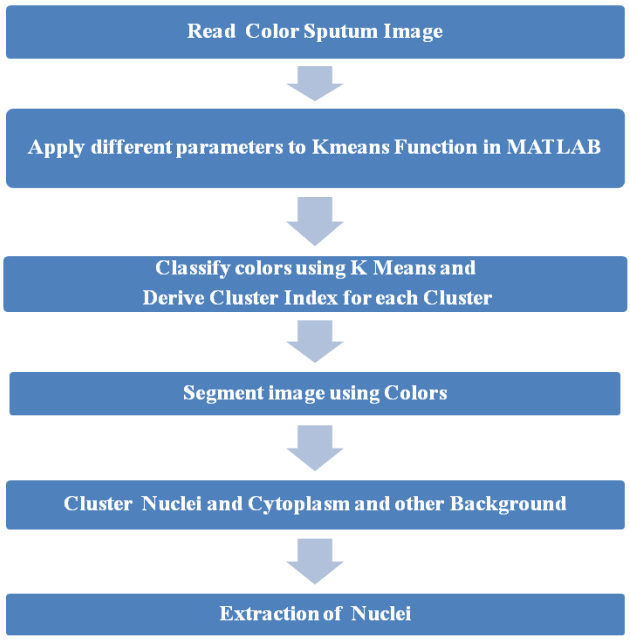


Figure 1 : Proposed Technique to segment the Sputum Image.

After reading sputum image in MATLAB, K Means algorithm is applied to this image and clustered the sputum image based on different color pixels.

Different parameters are applied to retrieve the best segmented results. K Means gives the best result of segmentation with Cosine Distance but it will take more time related to Euclidean and City Block. Main criteria of this proposed method is to obtain the best segmented result of sputum image in Nucleus, Cytoplasm and Debris Cells.

IV. RESULTS AND DISCUSSION

With the extracted Nuclei in result, we can derive the features of Nuclei to decide the given sputum image is cancerous or not.

Normal Sputum Image :

- Uniform shape and size of Nuclei
- No distortion in Nuclei

Cancerous Sputum Image:

- If extracted nuclei is not uniform in shape and size.
- If there is a distortion in size of Nuclei.

- If more than two nuclei in one Cytopla

A. Results.

- Input Sputum Image [8]:

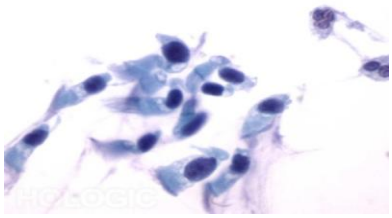


Figure 2 : Input Sputum Image

- Extracted Cytoplasm from given Sputum Image :

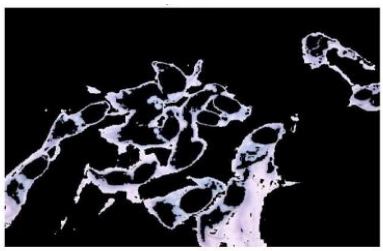


Figure 3: Extracted Cytoplasm

- Extracted Nuclei from Input Sputum Image :

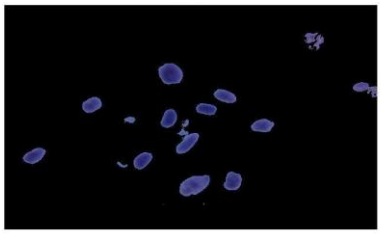


Figure 4: Extracted Nuclei

B. Comparison Between K Means and FCM :

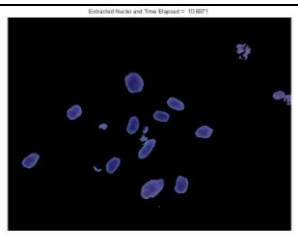
Extracted Nuclei (K Means)	Extracted Nuclei (FCM)
	

Figure 5: Extracted Nuclei of K Means and FCM

Methods	Time Taken	Output Image	Segmentation Results
FCM	27.228 seconds	Black & White	Improper , Noisy Nucleus
K Means	18.4325 seconds	Color Image	Well Segmented

TABLE 1 : Comparison Between K Means And Fcm Results.

As shown in Figure 4 .We can see the result difference of extracted nuclei of K Means and FCM.

Though FCM deals with gray scale image but it does not provide good segmented nucle.FCM provides noisy nuclei while K Means provide better and accurate segmentation result

V. CONCLUSION

In this paper we have seen different methods for Sputum image segmentation. Proposed technique of K Means for image segmentation works very well. The Proposed method can be used to segment the sputum image into Nuclei, Cytoplasm and Background. K Means segments the sputum image based on color pixels and it takes very less time compare to other methods. K Means segmentation results are more accurate and reliable than FCM. K Means succeeded in detecting and segmenting the nuclei and cytoplasm regions. FCM provides segmentation result in much more time than K Means and with noisy Nucleus. With the help of MATLAB tool and with K Means algorithm and feature extraction of Nuclei, Lung Cancer can be detected early using Sputum Images.

REFERENCES

[1] Cancer Care, <http://www.lungcancer.org/find_information/publications/163-lung_cancer_101/265-what_is_lung_cancer>

[2] Toni Johnson, The World Health Organization(WHO), < http://www.cfr.org/public-health-threats/world-health-organization-/p20003, 2011>

[3] American Cancer Society <http://www.cancer.org/acs/groups/cid/documents/webcontent/003115-pdf.pdf>.

[4] WebMD Medical Reference from Healthwise, <http://www.webmd.com/lung/sputum-cytology>

[5] Fatma Taher, Naoufel Werghi , Hussain Al-Ahmad and Christian Donner. "Extraction and Segmentation of Sputum Cells for Lung Cancer Early Diagnosis" ISSN 1999-4893,2013

[6] Sammouda , Fatma Taher, Naoufel Werghi, Hussain Al-Ahmad, Rachid "Lung Cancer Detection by Using Artificial Neural Network and Fuzzy Clustering Methods.", American Journal of Biomedical Engineering 2012,2(3):136-142

[7] Suman Tatiraju , Avi Mehta "Image Segmentation using k-means clustering, EM and Normalized Cuts" <http://www.ics.uci.edu/~dramanan/teaching/ics273a_winter08/projects/avim_report.pdf>

[8] Hologic Inc. Cytology Stuff <http://www.cytologystuff.com>.

