Extraction of Nuclear Region from Sputum Images through Pixel Classification for Early Lung Cancer Detection

Vineet Kumar
Department of Computer Science and Engineering
Lovely Professional University, Punjab, India

Hitesh Sharma
Department of Computer Science and Engineering
Lovely Professional University, Punjab, India

ABSTRACT
In today's world a very vast study has been made in the field of oncology. Now days, Lung Cancer constitutes the major portion of all deaths that occur due to cancer. This deadly disease Cancer is staged according to its severity and also up to where it has spread. It has been said by many doctors and researchers that a survival rate of five years can be increased if lung cancer is diagnosed in the early stages. There are many techniques available to diagnose the cancer. Two of them are Sputum cytology and Fine Needle Aspiration Cytology (FNAC). In these techniques, the presence of cancer cells in the cytological samples is examined. After that Normal cells are differentiated from Cancer cells on the basis of cell and nucleus appearance. In case of images a count of the number of pixels in nucleus, cytoplasm and background region can be made. In this paper the work on extraction of nucleus region from cytological sample images of sputum is presented. The reason is that, nucleus contains deoxyribonucleic (DNA), and DNA is responsible for cell formation and deformation. Threshold is used as a preprocessing step and problem of segmentation is viewed as a three class pixel classification problem. The classification technique used is Bayesian Classification.

Keywords
Sputum, Cytology, Biopsy, DNA, Lungs, Nucleus, Cytoplasm, HNN, FCM, FNAC, Debris, PAP staining.

1. INTRODUCTION
Along with heart, our chest contains two sponge-like organs called Lungs. In human body there exist two lungs i.e. left lung and right lung. Every lung is divided into lobes. Right lung is divided into three lobes and left lung is divided into two lobes. Lung structure starts from trachea. Trachea is further divided into tubes called as bronchi. Bronchi further develop as Primary Bronchi, Secondary Bronchi and then Tertiary Bronchi. Going down further, bronchi develops as bronchioles. Then there appear tiny air sacs called alveoli. These air sacs are responsible for inhaling and exhaling of oxygen and carbon-dioxide respectively.

Cancer is named according to where it has started. Lung cancer can start in any region of lung. It could be the cells lining the bronchi and in any part of the lung such as bronchioles and alveoli. It is known that cell is a basic unit of life and the basic unit of life of cancer is also a cell. Because cell contains nucleus in which the deformation starts, so in this work extraction of the sputum cell nucleus is presented. This sputum is a mixture of mucus and saliva. Mucus comes up through trachea whenever lungs suffer from any infection or allergy and is brought up via a hair like structure called cilia. It then mixes up with saliva which is secreted by salivary glands present in mouth. This sputum is coughed up and taken as sample. These samples are then stained with blue or red dyes and studied under microscope for the presence of cancer cells. In normal cases, a cell maintains a uniform shape and contains a single nucleus. But when DNA gets damaged, the cell either repairs the damage or it dies. But in case of cancer cells the damaged DNA is not repaired and instead of dying, cell goes on making new cells that body does not need. Then these cells multiply and forms tumour. In very early stages the cancer remains confined to the region where it has started. But soon it metastasizes.

The other imaging techniques like X-Ray, CT-Scan are also available for the detection of cancer. But here sputum cytology is used because it is economical to have regular tests of it. Sputum samples can be taken easily as a person normally coughs up. In comparison to other techniques, some people may be allergic to the fluid that is injected in body in case of CT scan. Also some may find difficulty to be in the tube like structure used for Magnetic Resonance Imaging (MRI). Both of these CT-Scan and MRI will only show the infected region and not the cells. Still it is a topic of discussion that which technique is helpful in decreasing mortality rate. FNAC can also be used for the purpose of detection in which a sample of mass is taken by a needle from the suspected region. These samples are then stained and studied under the microscope for the presence of cancer cells. Lung Cancer spreads to other regions of body very fast and most of the time it is detected when it has already been metastasized. Once it metastasizes to other regions of an organ or to other organs, it is very difficult to treat. The mortality rate can only be decreased when it is found early. So it is essential to find out the cancer cells at early stages.

For the purpose of finding cancer at very early stages, it is essential to extract the nucleus of the cell from the cytoplasm region. So in first step extraction of cell region from the background region was done because background region contains debris cells which are cells other than sputum cells. It is necessary to remove these cells for making better analysis. Then nucleus region was extracted for making analysis that whether cell is normal or abnormal. Because once nuclear region is found, malignancy or beginning stage of a cell can be determined.

Some techniques have already been used in image processing for this purpose. But in this paper cytological digital images of sputum are used. For removing the debris region, threshold was used. And for segmentation, Bayesian Classification was applied. Afterwards the accuracy of Bayesian classifier was measured with different color representations i.e. gray scale, RGB and HSV. Bayesian classifier was applied for classifying the pixels in three classes or three regions. These three regions

were named as Nucleus Region that will tell about the numbers of pixels present in the nuclear region, Cytoplasm Region that will tell about the numbers of pixels present in the cytoplasm and Background Region that will tell about the numbers of pixels present in the Background.

Because Bayesian classification is a supervised learning so, for classifying every pixel in its particular class, Bayesian classifier was trained with different feature vectors. These feature vectors were prepared manually. Feature vectors were prepared through different color spaces i.e. RGB colour space, HSV colour space and Gray colour space. Then classifier was trained with these feature vectors. Afterwards, accuracy of classifier with different feature vectors was measured. These trained classifiers were named as Gray Pixels Classifier that works with Feature Vector of Gray Values of pixels, RGB Pixels Classifier that works with Feature Vector of RGB Values of pixels and HSV Pixels Classifier that works with Feature Vector of HSV Values of pixels.

2. RELATED WORK

Various approaches have already been made to extract nucleus region from cytoplasm region for making better analysis about whether the cell is a cancer cell or a normal cell. Although there is advancement in oncology, but lung cancer remains the primary cause of cancer deaths worldwide. Because early detection of lung cancer increases the survival rate so doctors, pathologists and researchers are using various imaging techniques like X-ray, CT-Scan, FNA biopsy and Sputum Cytology to make diagnosis early and thus increase the survival rate. Sputum Cytology is an economical and non-invasive technique. Research has been made in this field to extract either the cell region or nucleus region.

Fuzzy k-c-means clustering algorithm for image segmentation was introduced in [1]. Here k-means and fuzzy-c-means clustering methods were combined to produce a more time efficient segmentation algorithm called as fuzzy-k-c-means clustering algorithm. They presented that thresholding which is the most basic technique for medical image segmentation, separates pixels in different classes depending on their gray level. They said that it approaches segmentation of scalar images by creating a binary partition of the intensity values of image. It finally determines an intensity value. This intensity value is called as threshold, which separates the desired classes. Classifier methods which were used for pattern recognition, partitions a feature space derived from the image using data with known labels. A feature space is a set of N*M matrix where N corresponds to the number of observations and M corresponds to the number of attributes. Classifiers are known as supervised methods since they require training data that are manually segmented and then used as for automatically segmenting new data.

A comparison between two methods was made in [2]. These methods are rule based method and Bayesian classicalism for the extraction of cell region from background and debris cell region, and after experimentation the Bayesian classicalism was found appropriate for classification of sputum cell region from background region. But they did not extract the nucleus region from cytoplasm region with this technique.

In [3] two more segmentation methods were used. These were Hopfield Neural Network (HNN), and Fuzzy C-Mean (FCM) clustering algorithm. They found that the HNN provides better, accurate and reliable segmentation results than FCM clustering in all cases. The HNN also segmented the nuclei and cytoplasm regions. FCM failed in the detection of the nuclei. FCM only detected a part of the nucleus not the complete nucleus in a particular cell. Also the FCM was not found sensitive to intensity variations because the segmentation error at convergence was found larger with FCM in comparison to HNN. According to the most recent estimates of the statistics provided by world health organization indicates that around 7.6 million deaths worldwide each year because of this type of cancer [3]. Furthermore, they found that mortality from cancer are expected to rise continuously, and will come near to 17 million worldwide in 2030. So, better methods are needed to extract the nucleus region for very early detection. A magazine in [4] provided us the knowledge about current trends in medical image analysis.

In [5] first images were enhanced through Gabor filter. It has given better results than other enhancement techniques. They only worked on the colored image enhancement and not extract the nucleus region and even not the cell region. In Features Extraction stage to obtain the general features of the enhanced and segmented image they used Binarization. A refined Charged Fluid Model (CFM) along with improved Otsu's method was used for the automatic segmentation of MRI images in [6]. This method gave better results than the previously used approaches.

In [7], a sober edge detection method was used which is based on finding the image gradient. It tells that intensity of the image will be maximum where there is a separation of two dissimilar regions thus an edge must exist there. So, on this basis they found the nodules in CT images. In [8], a new variation level set algorithm without re-initialization was used. They also used thresholding to reduce the noise component of the images.

In [9] glandular cells were detected using multiple colour spaces and two clustering algorithms. These clustering algorithms were K-means and Fuzzy C-means. In [10] an overview of whole process for processing digital images for lung cancer detection is given. This paper describes all the necessary steps required for the better performance starting from the pre-processing till the very end phase extraction of features.

3. METHODOLOGY

The work done in this paper includes the following steps:

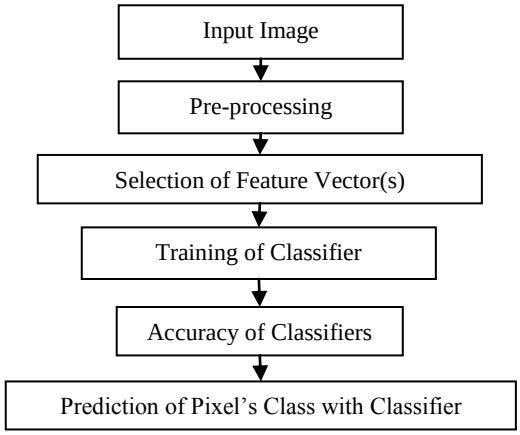

Figure 1: Workflow

3.1 Input Image

Digital images of stained sputum were obtained from Indira Gandhi Medical College (IGMC) Shimla which is a Regional Cancer Centre. Images of Sputum Slides stained with

Papanicolaou (PAP) smear were used. Slides can either be 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. Then these stained slides are kept under Microscope, Magnified and Images are taken by Digital Camera fitted with Microscope. A Sample of Sputum input image used as input is shown in Figure 2.

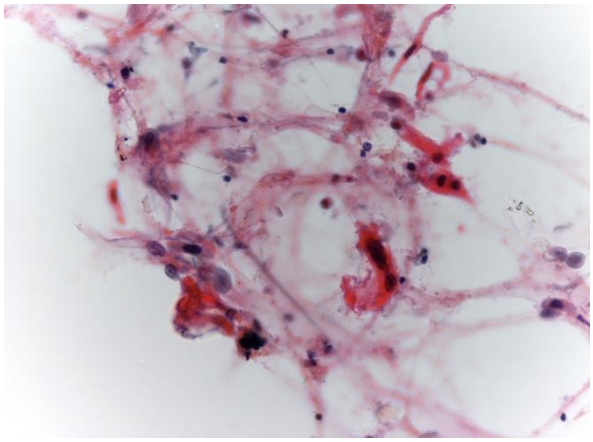

Figure 2: Sample Input Image

From Figure 2 the dark red sputum nucleus with clear red cytoplasm region can be seen. The blue debris cells are also present. So the image is not easy to analyse because it has some cluttered regions. So it is necessary to remove these debris cells first so that we are left only with cell region and then nucleus region can be detected with some ease.

3.2 Pre-processing

At first, it is necessary to Differentiate between Cell Region and Background Region. This background region contains Debris blue cells. These are cells other than Sputum Cells. For all this a Filtering algorithm was applied. Thresholding was used as a filtering algorithm. This thresholding depends on Staining Method used i.e. either Blue Dyes in which Sputum Nucleus becomes dark blue with clear blue cytoplasm and debris cell nucleus region becomes dark red with clear red cytoplasm region or Red Dyes (opposite to Blue Dye).

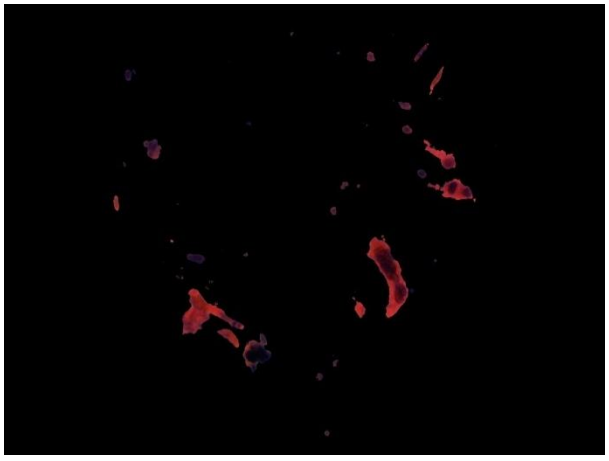

Figure 3: Threshold of Input Image

As much of the work has been done in the field of thresholding, the interface of ImageJ was used for this purpose and only the Blue component of RGB was varied between 0-255 by keeping the range of Red and Green region values full. The value of blue component was kept different for different images. After pre-processing, a sample of threshold image is as shown in Figure 3.

It can be seen in Figure 3 that now there are only three regions in the images i.e. sputum nucleus region, cytoplasm region and background region. Most of the debris region was removed in this step for making the analysis better.

3.3 Selection of Feature Vector(s)

Bayesian Classifier comes under supervised learning technique. In supervised learning the classifier is trained by providing it samples and targets. Samples are the input values provided to the classifier and target values are the desired output values respective to their input values. Each row of sample will contain one observation and each observation will contain a number of feature values. After training a classifier with these feature vectors new unseen values were passed to be predicted by the classifier. The preprocessed images were converted into gray, RGB and HSV color spaces and then features were extracted from these images manually by looking at their Gray values, RGB values and HSV values respectively. Then these features were converted into N*1 vector in case of gray images and N*3 in case of RGB and HSV images.

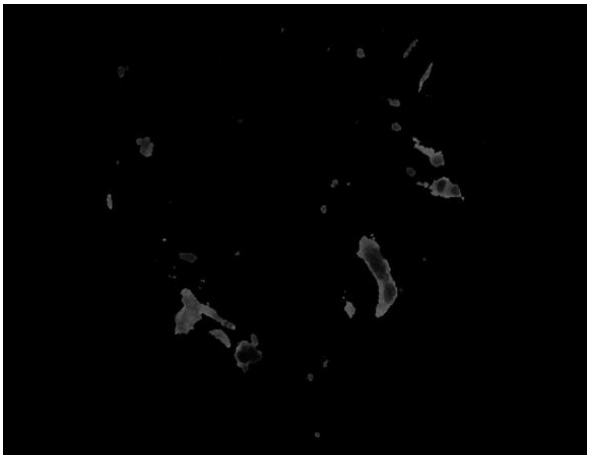

Figure 4: pre-processed sample gray scale image

A pre-processed sample gray scale image is shown in Figure 4 and gray values feature vector obtained from such images is shown in Table 1.

Table 1: Gray values Data Set

Gray Value of a Pixel	Target Value
60	Cytoplasm
70	Cytoplasm
80	Cytoplasm
65	Cytoplasm
75	Cytoplasm
65	Cytoplasm
85	Cytoplasm
90	Cytoplasm
10	Nucleus
15	Nucleus
20	Nucleus
25	Nucleus
30	Nucleus
35	Nucleus
40	Nucleus
45	Nucleus
5	Nucleus
0	Background

The pre-processed sample HSV image is shown in Figure 5.

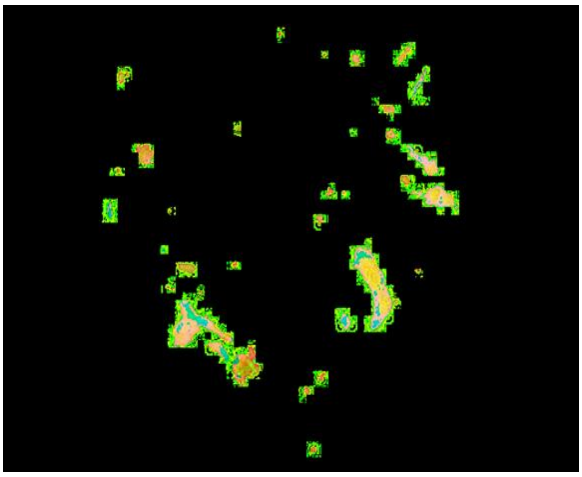

Figure 5: Sample HSV image

The feature vector obtained from such images is shown in Tables 2. The first three columns of feature vector are Hue, Saturation and Value. These are input and fourth column is the target class respective to input values.

Table 2: HSV Values Dataset

H Value of a pixel	S Value of a pixel	V Value of a pixel	Target Class
0	0	0	Background
0.33	1	0.01	Background
0.5	1	0.01	Background
0.33	1	0.02	Background
0.75	1	0.01	Background
0.9	0.66	0.21	Nucleus
0.88	0.68	0.21	Nucleus
0.91	0.63	0.2	Nucleus
0.89	0.63	0.22	Nucleus
0.93	0.82	0.2	Nucleus
0.98	0.74	0.55	Cytoplasm
0.99	0.7	0.53	Cytoplasm
0.96	0.67	0.49	Cytoplasm

Similarly the feature vector of 60*3 dimensions was obtained from different pre-processed RGB images and then further processing was done.

3.4 Training of Classifier

The main objective of this paper is to extract the nucleus region from the cytoplasm region, so the classifier that was used earlier was extended from two classes to three classes. These three classes are Background region, Cytoplasm Region and Nucleus region. Bayesian Classifier will calculate the conditional probability of having a pixel in Nucleus class, Cytoplasm class and Background class. It is based on Bayes rule, which says that

$$p(cj \mid d) = p(d \mid cj)p(cj)/p(d)) \quad (1)$$

Here

p (cj | d) is the probability of instance d being in class cj, This is what we are trying to compute

p (d | cj) is the probability of generating instance d given class cj,

p (cj) is the probability of occurrence of class cj, This is just how frequent the class cj, is in our database

p (d) is the probability of instance d occurring. This can actually be ignored, since it is same for all classes

So Bayesian Classifier works well when we have only one feature in our feature vector, because in case it doesn't assume attributes to have independent distributions, which is an essential condition of Bayesian Classifier. So, to simplify the task, naïve Bayesian classifiers assume attributes to have independent distributions, and thereby estimate

$$p(d \mid cj) = p(d1 \mid cj) * p(d2 \mid cj) * \ldots * p(dn \mid cj) \quad (2)$$

Where

p (d | cj) is the probability of class cj generating instance d, which is equals to

p (d1 | cj) the probability of class cj generating the observed value for feature 1, multiplied by p (d2 | cj).

p (d2 | cj) the probability of class cj generating the observed value for feature 2, multiplied by p (dn | cj).

This method was used to make the objects of NaiveBayes Classifier by giving input the samples defined as feature vectors in tables defined above. Then this classifier can be used to predict the classes of the new samples.

3.5 Accuracy of Classifiers

Now the classifiers are ready and can be used to predict the class of new pixel values. To classify Gray Scale values, RGB values and HSV values three classifiers were made by training them with their respective samples and targets. Now for better analysis, it is necessary to find out which classifier is giving less misclassification. For testing the classifiers, the input samples that were used as input were passed to the classifier

and their corresponding output values were checked. So, there were three test samples for checking the accuracy of classifier based on gray values, RGB values and HSV values. This misclassification error was calculated for every classifier. To calculate misclassification the method used is as follows:

$$\text{Misclassification error} = \text{Pixels Correctly Classified} / \text{Total Number of Pixels Presented to the Classifier} \quad (3)$$

Misclassification error in case of all three pixel values classifiers are presented in Table 3. By observing the results in table 3, it was found that misclassification in case of HSV feature vector is absolute zero. So, further observations were made on the basis of these results.

Table 3: Misclassification in case of Three Classifiers

	Misclassification Error
Gray Pixel Values Classifier	0.07%
RGB Pixel Values Classifier	0.05%
HSV Pixel Values Classifier	0.0%

3.6 Prediction of Pixel's Class with Classifier

Now the classifiers are ready to classify the pixels of new unseen samples values or all pixels of the image. For this first the image was read into the workspace. Then all three classifier's predictions were seen. To classify the pixels with a Gray Scale classifier, the image was converted to gray color space. Then a column vector of gray values of every pixel of this converted image was obtained by reshaping the gray scale matrix obtained. When RGB image was converted to gray scale image and a matrix was obtained in gray space, we have got a 1228800x1 column vector. Now this vector was passed as an argument to the classifier. After experimentation, number of pixels classified for sample image in Figure 4 in different regions is as follows:

Table 4: Results of Gray classifier for Figure 4

Class	Background	Nucleus
Number of pixels	1202469	26331

From the results present in Table 4 it is clear that Gray pixel values were not classified in all three classes. It is due to misclassification error as shown in table 3. Further, to classify the pixels with a RGB classifier, the image was converted to RGB color space. Then a vector of RGB values of every pixel of this converted image was obtained by reshaping the RGB matrix obtained. In this case a 1228800x3 vector was obtained, where 3 specify three values of pixels i.e. R, G and B. Now this vector was passed as an argument to the classifier. Due to misclassification of 0.05%, the results of RGB classifier were also not found satisfactory.

So, only HSV classifier was used which is giving a misclassification of 0.00% in this case. The results of classifier with HSV values for sample in Figure 5 are shown in Table 5.

Table 5: Results of HSV classifier for Figure 5

Class	Background	Cytoplasm	Nucleus
Number of pixels	1205080	13934	9786

So Bayesian Classifier with HSV feature Vector is able to classify pixels in all three classes.

4. CONCLUSION AND FUTURE WORK

It is clear from the findings that Color Representation has a major impact on the classification come segmentation. Bayesian Classifier with Gray and RGB feature vectors is not able to classify all the pixels correctly. But Bayesian classifier with HSV values feature vector is able to classify pixels in all three specified classes i.e. Nucleus Region, Cytoplasm Region and Background Region and also with a misclassification of zero. Further some post-processing technique can be applied on the classified pixels to visualize the results in the form of images i.e. assigning a different label to different class.

5. REFERENCES

[1] Ajala Funmilola A, Oke O.A, Adedeji T.O, Alade O.M, Oyo Adewusi E.A, "Fuzzy k-c-means Clustering Algorithm for Medical Image Segmentation", Journal of Information Engineering and Applications, ISSN 2224-5782 (print) ISSN 2225-0506 (online), Vol 2, No.6, 2012

[2] Christian D., Naoufel W., Fatma T., Hussain, "Cell Extraction from Sputum Images for Early lung Cancer Detection", IEEE 978-1-4673-0784-0/12, 2012

[3] Fatma T., Naoufel W., Hussain, Rachid S., "Lung Cancer Detection by Using Artificial Neural Network and Fuzzy Clustering Methods", American Journal of Biomedical Engineering, 136-142 DOI: 0.5923/j.ajbe.20120203.08, 2012

[4] "Medical Image Analysis", IEEE Pulse, 2154-2287/11/2011

[5] Mokhled S. AL-TARAWNEH, "Lung Cancer Detection Using Image Processing Techniques", Leonardo Electronic Journal of Practices and Technologies, ISSN 1583-1078, Issue 20, January-June 2012

[6] Nagesh V., Srinivas Y., Suvarna Kumar G, Vamsee Krishna V, "An Improved Medical Image Segmentation Using Charged fluid Model", International Journal of Engineering and Applications (IJERA) ISSN: 2248-9622, Vol. 2, Issue 2, pp.666-668, Mar-Apr 2012

[7] Nikita P., Sayani N., "A Novel Approach of Cancerous Cells Detection from Lungs CT Scan Images", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN 2277 128X, Volume 2, Issue 8, August2012

[8] Parsh Chandra B., Md. Sipon M., Bikash Chandra S. and Mst. Tiasa K., "MRI Image Segmentation Using Level Set Method and Implement a Medical Diagnosis System", Computer Science & Engineering: An International Journal (CSEIJ), Vol. 1, No. 5, December 2011

[9] Sajith Kecheril S, D Venkataraman, J Suganthi and K Sujathan, "Segmentation of Lung Glandular Cells using Multiple Color Spaces", International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.2, No.3, June 2012

[10] Sonit Sukhraj Singh, Anita Chaudhary "Lung Cancer Detection using Digital Image Processing", IJREAS Volume 2, Issue 2 ISSN: 2249-3905, (February 2012)