Automatic Sputum Color Image Segmentation for Lung Cancer Diagnosis

Fatma Taher, Naoufel Werghi and Hussain Al-Ahmad
Department of Electronic and Computer Engineering, Khalfan University
UAE
[e-mail: fatma.taher, naoufel.werghi, alahmad]@kustar.ac.ae]

Received August 6, 2012; revised November 14, 2012; accepted December 6, 2012;
published January 29, 2013

Abstract

Lung cancer is considered to be the leading cause of cancer death worldwide. A technique commonly used consists of analyzing sputum images for detecting lung cancer cells. However, the analysis of sputum is time consuming and requires highly trained personnel to avoid errors. The manual screening of sputum samples has to be improved by using image processing techniques. In this paper we present a Computer Aided Diagnosis (CAD) system for early detection and diagnosis of lung cancer based on the analysis of the sputum color image with the aim to attain a high accuracy rate and to reduce the time consumed to analyze such sputum samples. In order to form general diagnostic rules, we present a framework for segmentation and extraction of sputum cells in sputum images using respectively, a Bayesian classification method followed by region detection and feature extraction techniques to determine the shape of the nuclei inside the sputum cells. The final results will be used for a (CAD) system for early detection of lung cancer. We analyzed the performance of a Bayesian classification with respect to the color space representation and quantification. Our methods were validated via a series of experimentation conducted with a data set of 100 images. Our evaluation criteria were based on sensitivity, specificity and accuracy.

Keywords: Lung cancer, CAD system, Bayesian classification, Region detection, Feature extraction.

http://dx.doi.org/10.3837/tiis.2013.01.005

1. Introduction

Lung cancer, especially the malignant type is one of the deadliest cancers. Over the last few years the incidence of malignant tumor has continuously increased, because the cure of the cancer depends highly on its early diagnosis followed by an appropriate surgical excision. Physicians use several techniques to diagnose lung cancer, such as chest radiograph and sputum cytological examination where a sputum sample can be analyzed for the presence of cancerous cells [1]. The most recent statistics according to the American Cancer Society estimate that 226,160 new cases will be diagnosed (116,470 in men and 109,690 in women) in the U.S and there will be an estimated 160,340 mortalities from lung cancer (87,750 in men and 72,590 among women), in 2012 [2]. Furthermore, the statistics from the World Health Organization (WHO), indicate that deaths caused by cancer will reach about 12 million people in 2030 [3].

Manual screening for sputum cells identification involves a labor-intensive task with a high false negative rate. Automatic screening will bring several advantages, such as improving the sensitivity of the test and a better accuracy in diagnosis by increasing the number of images that can be analyzed by computer. Recently, medical researchers have proven that the analysis of sputum cells can lead to a successful diagnosis of lung cancer [4]. For this reason, we attempt to develop an automatic diagnostic system for detecting lung cancer in its early stages based on the analysis of sputum color images.

Several research papers have addressed the segmentation of the sputum cells. There are many algorithms which have been proposed in other articles for medical image segmentation, such as histogram analysis, regional growth, edge detection and pixel classification. Good reference of these techniques can be found in [5]. Other authors have considered the use of color information as the key discriminating factor for cell segmentation for lung cancer diagnosis [6]. The analysis of sputum images has been used in [7] for detecting tuberculosis consisting of analyzing sputum images for detecting bacilli shape. A survey for extracting different features such as, shape and size of micro-organism containing a few morphtypes from digital images can be found in [8].

The detection of lung cancer by using sputum color images was introduced in [9] where the authors presented unsupervised classification technique based on Hopfield Neural Network (HNN) to segment the sputum cells into cancer and non cancer cells to be used in the diagnosis process. They used an energy function with cost term to increase the accuracy in the segmented regions. The HNN successes in segmenting the sputum color image cells into nuclei, cytoplasm and clear background classes. Furthermore, it can make a crisp classification of the cells after removing all debris cells. However, the method has limitations due to the problem of early local minimum of HNN. The disadvantage is in the overlapping cells which are considered as one cluster.

The authors in [10] completed the work which has been done in [9], using an automatic computer aided diagnosis (CAD) system for early detection of lung cancer based on the analysis of pathological sputum color images. The RGB color space was used to represent the color images. Two segmentation processes have been used: the first one was Fuzzy C-Mean Clustering algorithm (FCM), and the second one was the improved version of the Hopfield Neural Network for the classification of the sputum images into background, nuclei and cytoplasm. These two latter regions were used as a main feature to diagnose each extracted cell. It was found that the HNN segmentation results are more accurate and reliable than FCM

clustering in all cases. However, the CAD system was associated with a high number of false positive rates, which make the chance of the patient's survival very low. Also, the constraint of fixing the numbers of clusters by the operators might in some cases compromise the quality of the segmentation. In addition to that, the main problem that faces any CAD system for early diagnosis of lung cancer is associated with the ability of the CAD system to discriminate between normal and abnormal cells (cancerous cells), therefore, we have to be very careful in choosing these features, because the success of the diagnostic system will be highly dependent on the extracted features.

With respect to the previous approaches, the novelty of our contribution is two-fold. 1) The detection and segmentation of sputum cell images using a Bayesian classification framework. In addition to its robustness, this framework allows a systematic setting of the classification parameter. We analyzed the performance of this method with respect to the color space representation and quantification. This approach will be described in section 2.

2) The region detection and feature extraction methods to formulate a rule in the CAD system: this part will be elaborated in section 3 and section 4, respectively. In section 5, the analysis phase is presented. In section 6 we describe and discuss the results. Finally in section 7 we present conclusion remarks and directions for future works.

2. Sputum Cell Segmentation

We used 100 images in this study provided by the Tokyo center of lung cancer in Japan. They all were stained according to the Papanicolaou standard staining method [11] (sputum reddish and background commonly blue) each of them had at least one sputum in it. The size was 768x512 pixels, provided in RGB space.

An example is shown in **Fig. 1** (a) where two sample images are presented. Each image contains one or more nucleus surrounded by cytoplasm. Moreover, there is an intensity variation in the cytoplasm region and the background of the image. The sputum cells are surrounded by many debris cells in the background of the images, thus forcing us to think about pre processing techniques which can mask all these debris cells and keep the nuclei and cytoplasm.

Furthermore, for each image a mask is manually made as ground truth data, dividing the images into sputum and non sputum segments as shown in **Fig. 1** (b). The aim of the segmentation process is to determine whether or not a pixel in the sputum image belongs to the sputum cell using its color information. The staining method allows to some level the sputum cell to have a distinctive chromatic appearance. Therefore, it is possible to separate automatically the sputum cell from the background using color attributes.

2.1 Bayesian Classifier with the Histogram Technique

In this approach, we address the cell segmentation problem using a probabilistic method which is based on the standard binary classification called Bayesian classification [12] due to its robustness in the training and testing process because we know the number of classes in this application either sputum cells or background cells, in addition to that this approach allows a systematic and methodologist estimation of the threshold parameters rather than using a heuristic rule based on trial and error testing. In Bayesian classification, let's assume x is the position in color space of the pixel we are looking at. *Class1* corresponds to the event where we have a cell pixel and *Class2* where we do not have a cell pixel. So in general we can say a

pixel part of a cell, *If* $p(Class1|x) \geq p(Class2|x)$, where *x* is part of a cell. From the application of the Bayesian theorem it follows that:

$$\frac{p(x|Class1)p(Class1)}{p(x)} \geq \frac{p(x|Class2)p(Class2)}{p(x)} \tag{1}$$

Since *p(x)* is equivalent in both sides, we get:

$$p(x|Class1)p(Class1) \geq p(x|Class2)p(Class2) \tag{2}$$

Now we introduce the parameter μij that describes the loss we get if *Class j* is selected, when the pixel is actually *Class i*, so the function for the loss we get when *Class1* is selected even though it is *Class2* is:

$$R_1 = \mu_{11}p(Class1|x) + \mu_{21}p(Class2|x) \tag{3}$$

And for the case that *Class2* is selected instead of *Class1* it is:

$$R_2 = \mu_{12}p(Class1|x) + \mu_{22}p(Class2|x) \tag{4}$$

Since $\mu 11 = \mu 22$ then:

$R_1 = \mu_{21}p(Class2|x)$ and $R_2 = \mu_{12}p(Class1|x)$, So R_1 is the loss that we expect when a sputum pixel is classified as a non sputum pixel, and R_2 is the loss when a non sputum pixel is selected as sputum pixel. In our work, we found that $R_1 < R_2$. So we get:

$$\frac{\mu_{21}}{\mu_{12}} < \frac{p(Class1|x)}{p(Class2|x)} \tag{5}$$

Using Bayesian theorem we get

$$\frac{\mu_{21}}{\mu_{12}} < \frac{p(x|Class1)p(Class1)}{p(x|Class2)p(Class2)} \tag{6}$$

After doing small transformation, we get the likelihood ratio on the right side:

$$\frac{\mu_{21}}{\mu_{12}}\frac{p(Class2)}{p(Class1)} < \frac{p(x|Class1)}{p(x|Class2)} \tag{7}$$

Since only 20% of all pixels in our database images belong to *Class1* we get:

$$4*\frac{\mu_{21}}{\mu_{12}} < \frac{p(x|Class1)}{p(x|Class2)} \tag{8}$$

The setting of the ratio $\lambda = \frac{\mu_{21}}{\mu_{12}}$ is based on the following reasoning.

In the context of cancer cell detection, cancerous cells usually determine when the size of the nucleus is larger than the size of the cytoplasm. In addition to that, false positives usually prevail over false negatives. Thus, the cytoplasm region will be increased if we mistakenly select a background pixel as a sputum pixel. This disproportionate the nucleus, and therefore, increases the likelihood of assessing a cancerous cell.

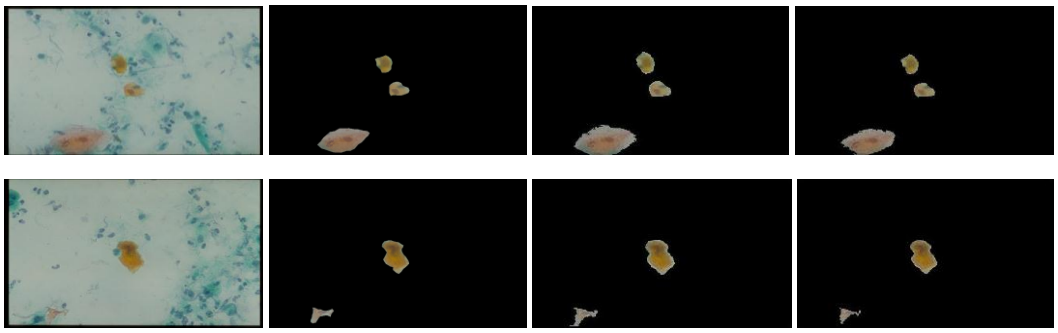
For this reason, the loss acquired in a false sputum cell classification should be assigned a larger weight than its counterpart in the opposite case, (e.g. loss acquired if the background class has been selected instead of the sputum), therefore, the ratio $\lambda = \frac{\mu_{21}}{\mu_{12}}$ should be set larger than 1.

The class condition pdfs *p(x|Class1)* and *p(x|Class2)* were estimated using histogram technique [13]. This approach is motivated by several reasons:

First, with the histogram technique, there is no need to make any assumption about the shape of the sputum and background probability density functions. On the other hand, when a specific form of the class-conditional pdf is assumed, as is the case with the Gaussian density models, some color spaces may prevail over others. Second, with the histogram technique, the Bayesian classifier can be designed very rapidly even with a large training set, as compared to other classifiers such as the Artificial Neural Network. Finally, in this application, the feature space has low dimensionality. The histograms were computed for different color spaces (RGB, YCbCr, HSV and L*a*b). Each channel in each color space is separated into bins (32, 64, 128 or 256). This was done for each combination of color space and number of bins.

For further explanations, we take RGB as color space and the number of bins is 32. Each channel in RGB color space is separated into 32 bins, and we get a 32x32x32 matrix, each dimension corresponding to one of the color channels (R, G or B). This was done twice so we could gain a histogram for the sputum pixel distribution and one for the non sputum pixels distribution. One visualization of the RGB histogram with 32 bins is shown in **Fig. 2**. As we can see in **Fig. 2**, the probability distribution of the pixels in RGB color space in 32 bins histogram. The upper row shows the distribution for the sputum pixels and the lower row shows the distribution for the non sputum pixels. To make the distribution more visible the 3D-histogram was projected down to 2 dimensions. In the first column the histograms were summed up over the first dimension (R), in the second column over the second (G), etc. Blue color means low probability and red color a higher one. Furthermore, the sputum pixels have a much larger variance in the RGB space than the non sputum pixels.

Fig. 1 (c) and **Fig. 1** (d) show two examples of sputum cell segmentation obtained with two different values of the ratio λ. **Fig. 1** (c) shows the output images from the Bayesian classification with λ= 2, and **Fig. 1** (d) shows the results with λ= 7. We can notice that the higher the threshold is, the smaller the size of the detected cell. The size of the detected cell for λ=7 is less than the size when λ =2, in addition to that, the nuclei and cytoplasm regions are detected correctly in **Fig. 1** (d) and are closer to the ground truth image in **Fig. 1** (b).



(a) (b) (c) (d)

Fig. 1. Samples of sputum cell segmentation results. (a) Raw images, (b) ground truth data, (c) cell detection with λ=2, and (d) cell detection with λ=7.

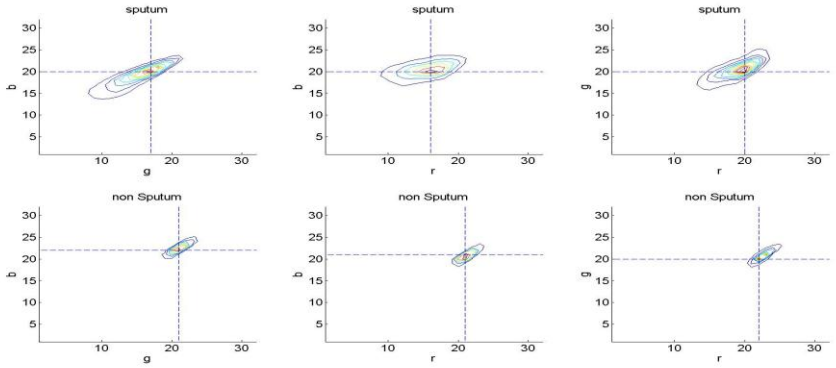


Fig. 2. Visualization of histogram (for RGB with 32 bins), the upper row shows the distribution of the sputum pixels, the lower row shows the distribution of the non sputum pixels.

3. Region Detection

The region-based approach is widely used in color image segmentation, because it considers the color information and spatial details at the same time. It creates regions by grouping together similar pixels with higher accuracy than pixel-based segmentation algorithms.

The idea of region detection is one of the most fundamental concepts used in image segmentation techniques. Several region detection strategies can be found in the literature [14]. In our case, the detection of the region by using the previous techniques is extremely difficult, and there are a lot of problems that need to be solved, since our extracted nucleus is considered as a binary image with two values 1 and 0, which represented the nuclei and background values, respectively. For this reason, we came up with our own new method to detect the connected regions in the nuclei cells in the sputum color images, where we used the concept of 8-connected components.

The algorithm starts searching for individual pixels (sometimes called seeds) and merges into the same region if their values are equal to one (this is our criterion). After which each tested pixel is compared to its immediate 8-neighboring pixels. If the criterion is fulfilled, then the tested pixel belongs to the same region, otherwise the tested pixel starts as a new region. The process of region detecting is continued until all pixels in the image belong to regions as

homogenous as possible, and by homogenous we mean that all pixels in a given region have the same value and are connected through their 8-neighbors. We used the connected-components labeling method to identify each region in the outlined image, through assigning a distinguished label to each region in a label matrix. Practically the region detection algorithm is described in **Fig. 3**:

```
Region_ Detection ( )
{
region_no = 1
for each pixel in the image
If (pixel_value= = 1and did not belong to any region yet)
{
Find_Region (pixel_index, region_no)
Region_no = region_no+1
}
Find_Region (pixel_index, region_no)
{
If(pixel_value= = 1and did not belong to any region yet)
label the pixel by region_no
call the Find_Region for each of its 8-neighbors
}
```

Fig. 3. Region Detection Algorithm

After that, an identification process must be performed to determine which regions have to be counted and which to be discarded as will be explained in the next section.

4. Feature Extraction

After detecting the nuclei regions, we extract prominent features that will help in efficient diagnosis of lung cancer at a very early stage. Feature extraction is a highly significant step in any CAD system. In the literature, we found among the features used in detecting the lung cancer, the following features [16]

1. Area of the nuclei region (A).
2. Area of the cytoplasm region (B) corresponding to nuclei A.
3. The Maximum Drawable Circle inside nuclei A (MDC).

Based on medical information the morphology, the size, and the growing correlation of the nuclei and its corresponding cytoplasm regions reflect the diagnostic situation of the cell life. We use the nuclei area (A), and the cytoplasm area (B) to compute the surface of the nuclei and cytoplasm, respectively, in order to obtain a ratio(R) between the two, where (R = A/B). In the normal case, the nuclei surface must be smaller than the cytoplasm surface. The third feature represents each nuclei cell by its corresponding maximum drawable circle (MDC).

The concept of MDC is similar to the concept of radius, however, our regions resemble a non convex shape thus, we cannot apply a radius to them. The MDC begins to draw a circle starting from a point inside a nuclei region, which must fulfill the condition that all pixels inside the circle belong to the region in the process. We tried all the pixels inside the region as

a starting drawing point. The process began by drawing a one pixel radius size circle inside the region. If the process succeeds, we increase the radius by one pixel and try to redraw the circle. We continue in this manner until we cover all pixels in a candidate region. Then we record the radius size of the pixel and compare it with the next drawing pixel in the same region, taking the pixel with the smallest value. We repeat the process to cover all candidate regions. At the end, each candidate region saves its maximum drawable circle to be used in the diagnostic process. We simulate the drawing of the circle inside the region process by examining the 8-neighbors of the starting point. If all neighbors belong to the same region as the starting point region, the drawing process succeeds, which means that a pixel radius size is achieved.

5. Analysis Phase

The main objective of the CAD system is to obtain a high level of true positives detection rate even for small cells and a low number of false negative. The term True Positives (TP's) describe the candidate cells that are classified by the CAD system as cancerous cells, which is true for these cells. The term False Positives (FP's) describe the candidate cells that are classified by the CAD system as cancerous cells, but in fact, are not cancer cells. The true cancer cells that are missed by the CAD are called False Negatives (FN's). The True Negatives (TN's) refer to the non-cancerous cells that are classified by the CAD system correctly.

To achieve this, a set of rule-based techniques have been used to preserve as much as possible the cancer cell regions, whereas at the same time, eliminating those with noncancerous cell regions that may exist among the segmentation results. Two diagnosis rules are constructed based on the medical history. All regions that have a MDC less than 5 pixels will be deleted, because these regions do not have a chance to form a tumor according to medical knowledge.

***Rule1*:**

1. *If* $(R < T_1)$ this cell is a normal cell and it is classified as a normal cell.
2. *If* $(T_1 \leq R < T_2)$ this cell is abnormal and it is classified as a candidate to be a cancer cell at level-1.
3. *If* $(T_2 \leq R < T_3)$ this cell is abnormal and it is classified as a candidate to be a cancer cell at level-2.
4. *If* $(R \geq T_3)$ this cell is abnormal and it is classified as a candidate to be a cancer cell at an advanced stage noted level-3.

(T_1, T_2 and T_3 are threshold values determined by experiments).

If $(R<<<T_1)$, in this case we cannot consider *Rule1* therefore, we cannot classify the cell as a normal or abnormal cell, and this happens due to the dispersion of the cytoplasm in the staining process. Special case in *Rule1* appears when more than one cell has connected cytoplasm regions, and in this case, *Rule1* will not be accurate to discriminate the normal from abnormal cells. For this reason, we apply the following morphological rule to the nuclei regions.

Rule2:

Compute the area of the maximum drawable circle (MDC) in the region (A) as follows:

Compute the nuclei area of the region (A) as follows:

Area (A) = The absolute value of the deviation (DV) between the two areas, Area (MCD) and Area (A) as follows:

If $(DV \geq T_4)$, Then this cell is a candidate at level-3.

Experimentally, from our database images, we found that the convenient threshold values are as follows: $T_1 = 0.75$, $T_2 = 0.85$, $T_3 = 0.95$ *and* $T_4 = 300$.

6. Experiments

In the first series of experiments, we conducted a comprehensive analysis for the effect of the Bayesian classifier on the sputum cells. The data partition we used for testing and training is depicted in **Table 1**.

Table 1. Sputum image data partition for training and testing

Dataset	# of Images	# of sputum pixels	# of background pixels
Training set	65	1,305,876	23,840,921
Test set	35	512,133	13,779,755

Quantitatively, the evaluation criteria are based on sensitivity, specificity and accuracy, in terms of TP, TN, FN and FP. The evaluation criteria are defined as follows [16]:

$$\text{Sensitivity} = \frac{TP}{(TP+FN)}$$

$$\text{Specificity} = \frac{TN}{(TN+FP)}$$

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

The sensitivity reflects the extent to which pixels are classified as sputum pixels are actually sputum pixels, specificity measures how close the background is classified, and the accuracy evaluates the overall correctly detected pixels. Furthermore, we used ROC curves for visualization of the performance. An ROC curve is the parametric curve which contains: False Detection Rate (t), Correct Detection Rate (t), where t is the classifier parameter equal to λ.

False Detection Rate is defined as the percentage of non sputum pixels that were classified as sputum pixels. True Detection Rate is defined as the percentage of sputum pixels that were classified as sputum pixels**.** We tested the Bayesian classifier in term of color representation and quantification on the sputum cell segmentation for different color spaces (RGB, YCbCr, HSV and L*a*b) with different histogram resolutions (32, 64, 128 or 256). The results are shown in **Fig. 4**. The ROC-curve for the four color representations for 256 histogram resolution is shown in **Fig. 4** (a), where we found that HSV space has the best performance across all the resolutions, followed by the RGB space. However, overall, the different color spaces show a close performance for resolutions above 64.

Fig. 4 (b) shows the ROC-curve of the RGB space for the different histogram resolutions. We clearly observe that the performance improves as the resolution increases, yet is remaining reasonable across the whole range. Similar behavior has been observed for the HSV space. On the other hand, we found that the performances of the YCbCr and L*a*b* spaces degrades

when the resolution drops below 64.

Fig. 4 (c) and **Fig. 4** (d) show the accuracy performance for the Bayesian classification in term of color representation and quantification for 128 and 256 histogram resolutions respectively. They show the accuracy criterion variation of the color spaces in function of the ratio λ. As we can see the accuracy improves as λ increases and reaches its maximum value within the range [15-17]. Moreover, we note that the accuracy improves as the histogram resolution increases. In **Table 2** we compile the best accuracy scored by the four color spaces for the different histogram resolutions. As can be seen, the HSV and RGB achieved the best accuracy over different histogram resolutions and they are very close to each other.

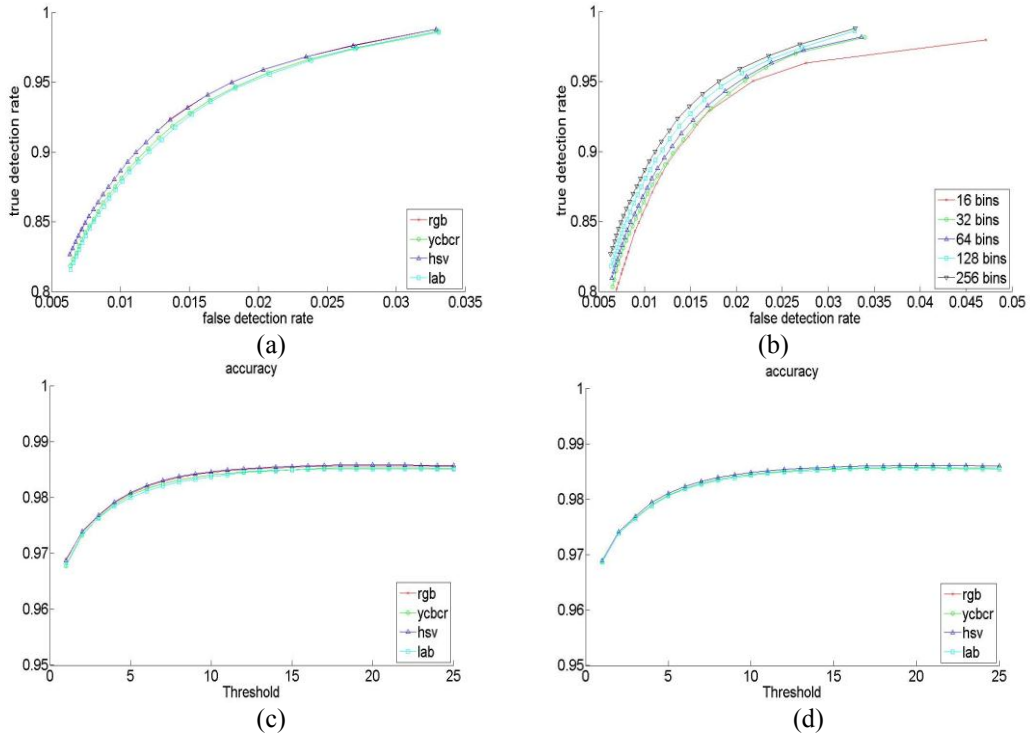


Fig. 4. (a) The ROC curve in four color spaces for a histogram resolution 256. (b) The ROC curve of RGB space for the different histogram resolutions. (c) The accuracy performance of the color spaces for a histogram resolution 128. (d) The accuracy performance of the color spaces for a histogram resolution 256

Table 2. Best Accuracy for each color space/quantization combination

Bins/Color spaces	RGB	YCbCr	HSV	L*a*b
16 bins	0.9843	0.9826	0.9848	0.9823
32 bins	0.9850	0.9838	0.9852	0.9820
64 bins	0.9853	0.9849	0.9855	0.9842
128 bins	0.9857	0.9853	0.9859	0.9851
256 bins	0.9861	0.9858	0.9861	0.9856

In the literature and to the best of our knowledge, we found that the only available technique for early lung cancer detection by using sputum color images is to utilize the threshold technique **[9, 17].** Based on that, **Table 3** represents a clear comparison between the

performances of the threshold method and the Bayesian classification. We found that the Bayesian classification achieved the best scores. It succeeded particularly in reducing the number of FN and improving the sensitivity. On the other hand, the specificity and accuracy are close to their counterparts in the threshold method which reveals a clear superiority of the Bayesian method.

Table 3. Performance of the Segmentation Process

Performance Measurements	Threshold Technique	Bayesian classification
Sensitivity	82%	89%
Specificity	99%	99%
Accuracy	98%	98%

In the second series of experiments, we tested the effect of region detection and feature extraction on the segmented sputum cells. We used the same evaluation criteria as in the sputum segmentation using the criteria: sensitivity, specificity and accuracy.

The pathologist-identified cells are used as the gold standard to analyze the accuracy of the proposed CAD system. We compared the results of 100 images in our database with the pathologist results, and among these images, 65 cancerous candidate cells are classified as cancerous cells by the proposed CAD system and the pathologist, 19 are classified as normal cells by both of them, and 16 are classified differently.

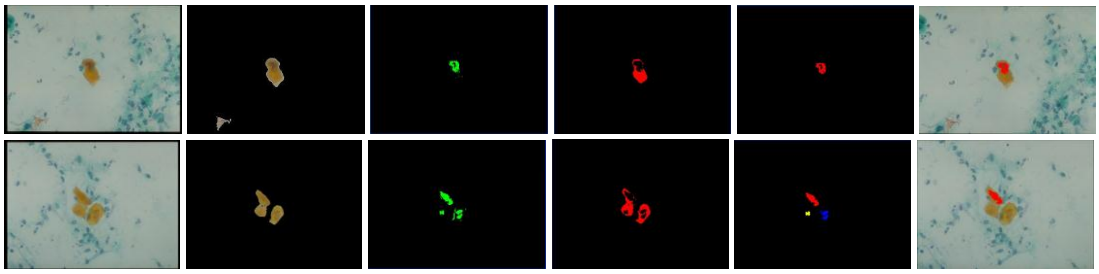
Table 4 shows the results of applying our CAD system to various numbers of images, and **Table 5** shows the performance of our CAD system. **Fig. 5** shows the results obtained with the first two sputum color images in **Table 4**, respectively. **Fig. 5** (a) shows the original raw images, **Fig. 5** (b) shows the segmented images, after applying the Bayesian classifier with λ =7, where the nuclei and cytoplasm are well extracted, and all the debris cells are removed. **Fig. 5** (c) and **Fig. 5** (d), show the nuclei and cytoplasm region extraction, respectively. **Fig. 5** (e) shows the region detection after applying the feature extraction process and after deleting the regions with a MDC less than 5 pixels, because these regions do not have a chance to form a tumor according to medical knowledge. Finally, **Fig. 5** (f) shows the affected part of the cell in the images. According to our system the sputum cell in the image which is located on the first row of **Fig. 5** (f) is a normal cell as represented in **Table 4.** The image on the second row of **Fig. 5** (f) contains the affected nucleus cell and this is a candidate with a cancer cell at level 1.

Table 4. Results after applying a feature extraction process

Images	Area (A)	Area (B)	R	DV	CAD/Diagnosis	Pathologist/Diagnosis
1	1128	2799	0.403	221	Normal	Normal
2	1337	1600	0.835	324	Cancer level-1	Cancer level-1
3	1524	1332	1.144	438	Cancer level-3	Cancer level-3
4	713	1127	0.632	90	Normal	Normal

Table 5. Performance of the CAD system

TP's	FN's	FP's	TN's	Sensitivity	Accuracy	Specificity
65	8	10	17	90%	92%	98%



(a) (b) (c) (d) (e) (f)

Fig. 5. CAD system results. (a) Original raw images, (b) cell segmented with Bayesian classifier λ =7. (c) Nuclei extraction, (d) cytoplasm extraction. (e) The regions detection in the nuclei cell. (f) The tested cells to be diagnosis, the cell in the first row is normal, and the cell in the secnd row is a cancerous cell.

Conclusion

In this paper, we present a framework for the detection and classification of lung cancer based on the analysis of sputum color images. Our methods are based on using a Bayesian classification with histogram techniques followed by a region detection and feature extraction processes. The Bayesian model allows the extraction of the nuclei and cytoplasm regions successfully. In addition to that, the assessment of the Baysian classification with regards to color reveals a close performance across the different color space for histogram resolution above 64. Afterwards, we determine the connected regions in the nuclei and analyzed them to extract a set of features to be used in the diagnostic rules. These rules are formulated in the next step to discriminate between cancerous and non-cancerous cells in the sputum color images. In this work, 100 samples of sputum images were analyzed. We have tested our CAD system, by comparing its diagnosis case to the diagnosis of an experimented pathologist of the same data. At the current stage the CAD system produces a reasonable accuracy equal to 92 %, with sensitivity and specificity equal to 90% and 98%, respectively. Nevertheless, this performance can be further improved via further basic morphological processing on the segmented image. In the future, we plan to increase our features range to be more discriminate and to consider a support vector machine in the classification of the lung cancer cells as soon as a more extended dataset is available.

References

[1] R. Kemp, M. Daniel, B. Turic, “Detection of Lung Cancer by Automated Sputum Cytometry”, *Journal of Thoracic Oncology*, vol. 2, no. 11, pp. 993-1000, Nov. 2007. Article (CrossRef Link).
[2] American Cancer Society, http://www.cancer.org/Cancer/LungCancer-Non-SmallCell/DetailedGuide/non-small-cell-lung-cancer-key-statistics, 2012.
[3] Toni Johnson, The World Health Organization (WHO), http://www.cfr.org/public-health-threats/world-health-organization-/p20003, 2011.

[4] A. Sheila and T. Ried, "Interphase Cytogenetics of Sputum Cells for the Early Detection of Lung Carcinogenesis", *Journal of Cancer Prevention Research*, vol. 3, no. 4, pp. 416-419, March, 2010. Article (CrossRef Link).
[5] S. Raut, M. Raghuvanshi, R. Dharaskar and A. Rau, "Image Segmentation – A State-of-Art Survey for Prediction", in *Proc. of the International Conference on Advanced Computer Control (ICACC '09)*, pp. 420 - 424, India, 2009. Article (CrossRef Link).
[6] F. Taher and R. Sammouda, "Identification of Lung Cancer based on Shape and Color", in *Proc. of the 4th IEEE International Conference on Innovation in Information Technology*, pp.481-485, UAE, Nov. 2007. Article (CrossRef Link).
[7] M. G. Forero, F. Sroubek and G. Cristobal, "Identification of Tuberculosis Bacteria based on Shape and Color", *Journal of Real time imaging*, vol. 10, pp. 251-262, 2004. Article (CrossRef Link).
[8] J. Liu, F. Dazzo, O. Glagoleva, B. Yu and A. Jain, "A Computer-Aided System for the Image Analysis of Bacterial Morphotypes in Microbial Communities", *Journal of Microbial Ecology*, vol. 41, no. 3, pp 173-194, 2001. Article (CrossRef Link).
[9] R. Sammouda, N. Niki, H. Nishitani, S. Nakamura, and S. Mori, "Segmentation of Sputum Color Image for Lung Cancer Diagnosis based on Neural Network", *IEICE Transactions on Information and Systems*, vol. E81, no. 8, pp. 862-870, August, 1998. Article (CrossRef Link).
[10] F. Taher and R. Sammouda, "Morphology Analysis of Sputum Color Images for Early Lung Cancer Diagnosis", in *Proc. of the 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010)*, pp. 296-299, Malaysia, May. 2010. Article (CrossRef Link).
[11] Y. HIROO, "Usefulness of Papanicolaou Stain by Rehydration of Airdried Smears", *Journal of the Japanese Society of Clinical Cytology*, vol. 34, pp. 107-110, Japan, 2003.
[12] S. Phung, A. Bouzerdoum and D. Chai, "Skin Segmentation using Color Pixel Classification: Analysis and Comparison", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 27, no.1, pp. 148-154, 2005. Article (CrossRef Link).
[13] F. Calderero, F. Marques and A. Ortega, "Performance Evaluation of Probability Density Estimators for Unsupervised Information Theoretical Region Merging Histogram", in *Proc. of the 16th IEEE International Conference on Image Processing (ICIP)*, pp. 4397 – 4400, 2009. Article (CrossRef Link).
[14] C. Grigorescu, "Contour Detection based on Non-Classical Receptive Field Inhibition", *IEEE Transactions on Image Processing*, vol. 12, no. 7,pp. 729-739, 2003. Article (CrossRef Link).
[15] A. McWilliams, B. Lam, T. Sutedja, "Early Proximal Lung Cancer Diagnosis and Treatment", *European Respiratory Journal*, vol.33, no.3, pp.656-665, 2009. Article (CrossRef Link).
[16] Margaret H. Dunham, *Data Mining Introductory and Advanced Topics*, 1st edition, Prentice Hall New Jersey, 2003.
[17] F. Taher, Naoufel Werghi and Hussain Al-Ahmad, "A Thresholding Approach for Detection of Sputum Cell for Lung Cancer Early Diagnosis", in *Proc. of the IET Conference on Image Processing (IPR 2012)*, pp. 1-6, July 3-4, London, UK, 2012. Article (CrossRef Link).