Segmentation Of TB Bacilli in Ziehl-Neelsen Sputum Slide Images Using K-Means Clustering Technique

R. A. A. Raof, M. Y. Mashor, S. S. M. Noor
Universiti Malaysia Perlis, Pauh Putra, Perlis, Malaysia
Universiti Sains Malaysia, Kubang Kerian, Kelantan, Malaysia
e-mail: rafikha@unimap.edu.my, yusoff@unimap.edu.my, ssuraiya@kck.usm.my

Abstract

Image segmentation is the most crucial steps in determining the accuracy of a medical diagnosis system that is based on image processing procedures. Therefore, it is important to select a suitable image segmentation technique to obtain good results and hence providing optimum accuracy for the developed diagnostic system. In this research, image segmentation procedure using k-means clustering approach has been considered for differentiating between pixels that represent TB bacilli and pixels that represents sputum or background. This paper presents the technique used to separate the TB bacilli and its background from the Ziehl-Neelsen sputum slide images. k-means clustering have been applied to those images followed by several extra rules. The resulted images show encouraging results, which indicate that the proposed segmentation method is able to filter out the TB bacilli pixels from the background pixels.

Keywords— *K-Means Clustering, TB Diagnosis, Image Segmentation*

1. INTRODUCTION

Tuberculosis (TB) is one of the oldest diseases known to affect humans. At the beginning of the new millennium, despite efforts in the past decade to bring the problem under control, TB remains the most important infectious disease in the world. WHO [1] mentions in its report that globally, in 2010, there were 8.8 million TB occurrences, with 1.1 million deaths from TB among Human Immunodeficiency Virus (HIV)-negative people. On top of that, HIV-associated TB contributed of an additional 0.35 million deaths. 65% of the estimated number of incident cases in 2010 comes from 5.7 million notifications of new and recurrent cases of TB. TB is affecting mostly young adults in their most productive years. The vast majority of TB deaths are happening in the developing world, with more than half occurring in Asia, thus making TB as a disease of poverty. In 2010, the majority of the estimated number of TB cases occurred in Asia (59%) and Africa (26%). There are also cases found in the Region of Eastern Mediterranean (7%), the Region of Europe (5%) and the America Regions (3%). Five countries with the highest number of occurrence cases in 2010 were India (2.0–2.5 million), China (0.9–1.2 million), South Africa (0.40–0.59 million), Indonesia (0.37–0.54 million) and Pakistan (0.33–0.48 million). An estimated one quarter (26%) of all TB cases worldwide occurred in India along, with another 38% of the cases found in China and India [1].

TB still remains a major global health problem despite the availability of highly efficacious treatment for decades. In 1993, an estimated 7–8 million cases and 1.3–1.6 million deaths occurred each year. This situation forced the World Health Organization (WHO) to pronounce TB as a global public health emergency [2].

More than one third of the world's total population, which are approximately two billion people, are infected with TB. In their lifetime, one in every 10 of those people will turn out to be sick with active TB. Those who are living with HIV are having a higher risk. In order to achieve the target under the Millennium Development Goals (MDG), WHO is working with other agencies. WHO also aims to reach all patients through primary health care and systems. TB spreads through

the air and therefore is contagious. Each person with active TB can infect on average 10 to 15 people a year if it is not treated. There are two TB targets that are set for 2015 and the world is currently on the right schedule to accomplish them. The first target is the MDG that aims to reverse and halt global incidence (in comparison with 1990); and the second target is Stop TB Partnership that aims of halving deaths due to TB (also in comparison with 1990).

Research on new TB diagnostic tools has been accelerated over the last few years and as a result the diagnostic pipeline has been growing rapidly. There have been important breakthroughs in TB diagnostics following increased investments in TB research and development in the past decade. Light microscope is still used in most low and middle-income countries as it is cheaper and has lower maintenance costs as compared to fluorescence microscope. However, images captured using light microscope produced unclear separation between the bacilli and the background, which makes the detection process more difficult.

Advances in image processing algorithms, as well as computer hardware and software have also led to implementation of computer-aided system to TB diagnosis. The system aims to assist microbiologist in detection, either by improving the visual appearance of the tubercle bacilli for ease of scanning process [3][4] or detecting the bacilli automatically from a given specimen [5][6][7]. In the analysis of the objects in digital colour images it is essential that the objects of interest can be distinguished from the background. The techniques that are used to find the objects of interest are usually referred to as segmentation techniques; segmenting the foreground from background. At present, there is no technique of segmentation that can be applied universally to all images [8].

In the field of medical images, several systems have previously been reported to use this approach of image segmentation techniques in their project. Forero *et al.* [6][9] applied adaptive colour thresholding technique to the images that have been captured using fluorescence microscopy. Veropoulos *et al.* [10] used an identification method based on shape descriptors and neural network classifiers. Wilkinson [11] proposed a rapid multi resolution segmentation technique based on computing thresholds for different areas in a monochromatic image.

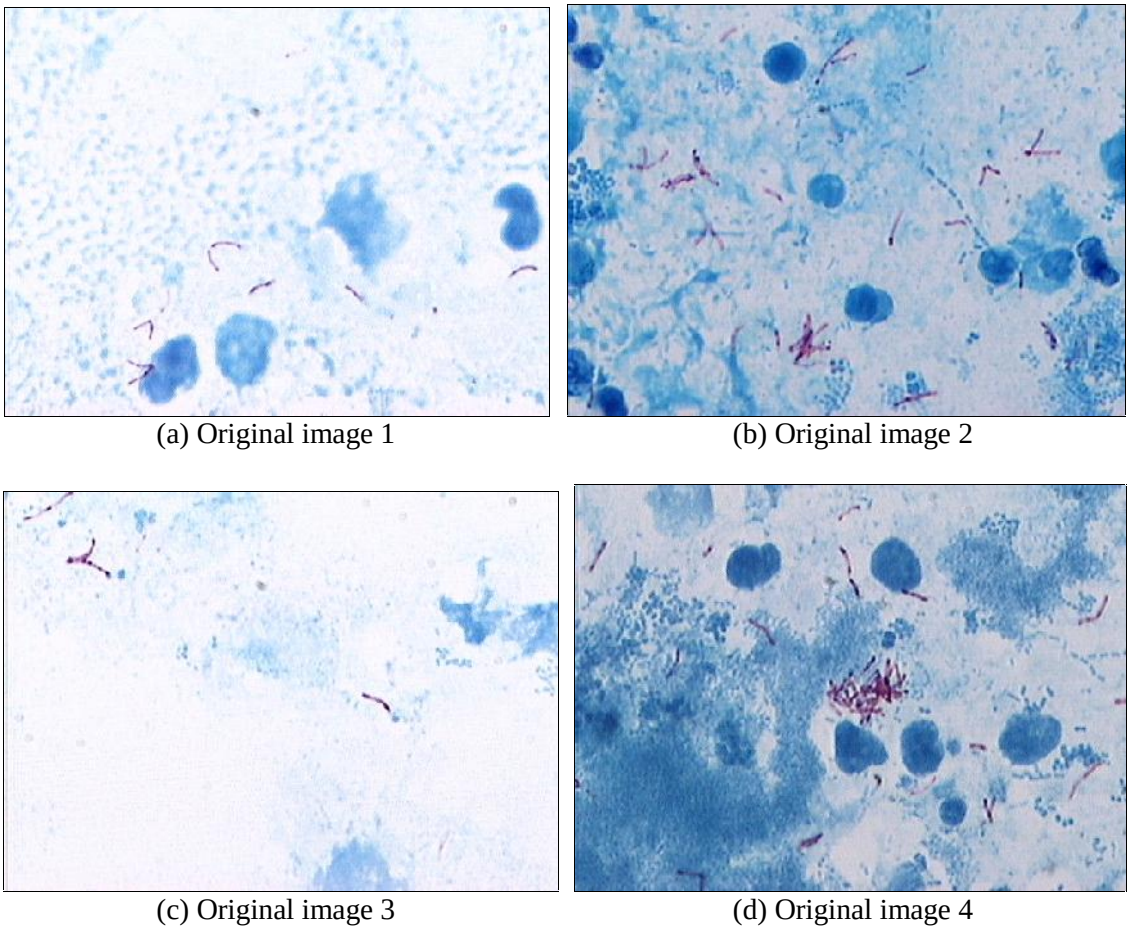
Image segmentation is the most crucial steps in determining the accuracy of a medical diagnosis system that is based on image processing procedures. Therefore, it is important to select a suitable image segmentation technique to obtain good results and hence providing optimum accuracy for the developed diagnostic system. However, it is preferable that the chosen segmentation technique to have a simple algorithm with the ability to perform good segmentation for the specified images.

The sputum specimen that has undergone the process of staining using Ziehl-Neelsen procedure will make the *mycobacterium* of TB appear red and other cells and organisms in the sputum smear sample will retain blue background as shown in Figure 1. Image segmentation is a part of image processing technique that will help to discriminate between the *mycobacterium* and sputum pixels in the digital image.

In this research, image segmentation procedure has been considered for differentiating between pixels that represent TB bacilli and pixels that represents sputum or background. This paper aims to present the technique used to separate the TB bacilli and its background from the Ziehl-Neelsen sputum slide images. This research used Ziehl-Neelsen sputum slide specimens obtained from the Department of Medical Microbiology and Parasitology, School of Medical Science, Universiti Sains Malaysia (USM), Kubang Kerian, Kelantan. Only sputum specimens from the patients are taken into account for the research. There is also TB diagnosis using tissue specimen, which is not included in this research.

2. METHODOLOGY

The sputum specimens consist of TB bacilli were obtained from Department of Medical Microbiology and Parasitology, School of Medical Science, Universiti Sains Malaysia (USM), Kubang Kerian. The sputum specimens have been stained using Ziehl-Neelsen staining procedure.



(a) Original image 1 (b) Original image 2

(c) Original image 3 (d) Original image 4

Figure 1 (a)-(d). Original sputum slide images consisting of TB bacilli

Figure 1 shows samples of images of Ziehl-Neelsen sputum specimens that have been captured automatically using an automated capturing system that has been developed to control the digital camera and the motorized stage of a light microscope. The original images are first saved into bitmap (*.bmp) files with the resolution of 800 x 600 pixels.

The problem faced in segmenting Ziehl-Neelsen sputum slide images is the inconsistency of the image brightness. Since the slides were prepared manually by microbiologist, there was possibility that the surface of the sputum slides were uneven. Uneven surface will produce variation in the image brightness when the images are captured from one end of the sputum slide to another. In this case, fixed-value threshold method might not be able to produce a good output image since the threshold value will only be suitable for images that are having standard brightness.

The segmentation process can be further divided into two steps: initial centres selection and k-means clustering. The k-means clustering algorithm has been used to segment the Ziehl-Neelsen sputum slide image into two clusters, which are TB and background (consist of sputum and other bacilli).

In order to proceed with the initial centre selection step, the value that will be used for clustering need to be decided first. The sputum slide images are colour images, therefore there are various possible value that can be used for clustering input. For example, each pixel in a colour image will have RGB values, HSI values or even the combination in between each individual value

that can be considered as the clustering input. Hence, study has been done to decide which element will be used as the input to the technique. The specified study has resulted that the ratio between red and green components can distinguished between TB bacilli pixels and background pixels. Therefore, this value has been adopted to be the input for the k-means clustering algorithm that is going to be implemented on the sputum slide images.

In order to view the important properties of each segment so that necessary features and accurate value of threshold can be obtained from the result, the information is being gathered in a table. In this table, among the features that are noted are the maximum, minimum and average values for each of the RGB components in TB bacilli and sputum respectively. The maximum, minimum and average values for each of the pixels are also noted to extract important characteristic of the RGB values that may be converted into threshold values. Summary of the findings from the study is visualized in Table 1.

Table 1. RGB information for Ziehl-Neelsen sputum slide images

	TB Bacilli			Background		
	MIN	MAX	AVERAGE	MIN	MAX	AVERAGE
RED	21	255	196	26	255	241
GREEN	18	254	190	37	255	245
BLUE	44	255	217	67	255	220

The minimum and maximum values for both TB bacilli and background pixels were almost the same. However, the average values for both categories were significantly different. Therefore, the average values were considered as the initial centres for k-means clustering. Since the main objective is to segment pixels in the image into two clusters, hence two centres are needed. represents centre for TB region and represents centre for background region. Therefore, centre for TB cluster, is taken using the RGB component (185, 161, 206), while centre for background cluster, is taken using the RGB component (235, 246, 254). Since the ratio between red and green components can distinguish between TB bacilli and background pixels, this value will be the input for the k-means clustering algorithm. Equation (1) shows the input for k-means clustering algorithm.

$$in(x,y) = \frac{red(x,y)}{green(x,y)} \tag{1}$$

where $red(x,y)$ and $green(x,y)$ are the original pixel values for red and green components, while $in(x,y)$ is the input for *k*-means clustering algorithm.

k-means clustering have been widely used for automatic image segmentation. For a colour image represented by RGB values, each pixel will be treated as an object. Each pixel has 3-dimensional vectors, which are R, G and B component respectively. However, in this research it has been reduced to 1-dimensional vector, whereby only the ratio of red and green components is used as the input. The total number of pixels (n) will depends on the size of the image. For a colour image, a non-adaptive *k*-means clustering may be implemented as follows:

i) Choose *k* pixels to be the initial centres of *k* clusters (C_k). In this study, two centres are chosen using the method previously mentioned.

ii) Calculate the Euclidean distance (E) between a pixel and each centre using Equation (2) by considering the colour information of the pixel. Assign the pixel to the cluster that has the nearest centre.

$$E = \left(\left\| in(x,y) - C_k \right\| \right)^2 \tag{2}$$

where $in(x, y)$ is the value of the input pixel and C_k is k^{th} cluster centre.

iii) Repeat step (ii) for all pixel in the image. Note that after this step, each pixel will be a member of one and only one cluster.

iv) When all pixels have been assigned, recalculate the positions of the *k* centres using Equation (3), by using the value of $in\,(x, y)$ calculated using Equation (1).

$$C_k = \frac{\sum in(x,y)}{n_k} \tag{3}$$

where C_k is k^{th} cluster centre, n_k is the number of pixels belonging to centre C_k and $in(x, y)$ is the values of input pixels belonging to the centre C_k.

v) Repeat Step (ii) - (iv) until the centres does not significantly move. Movement of the centres should be approximately less than 5%.

After the procedure has been performed, pixels that belong to the cluster associated with the background is deleted, or changed to white colour. On the other hand, pixels that are associated with TB bacilli will retain its original colour.

Apart from the procedure of *k*-means clustering, a few general rules have also been considered to ensure the optimum accuracy of the segmentation procedure. These rules are applied after the *k*-means clustering procedure. The general rules that are included in the procedure are shown in Equation (4) and (5).

$$out\,(x,y) = \begin{cases} white, & blue\,(x,y) < green\,(x,y) \\ original, & blue\,(x,y) \geq green\,(x,y) \end{cases} \tag{4}$$

$$out\,(x,y) = \begin{cases} white, & blue\,(x,y) > max\_blue \\ red, & blue\,(x,y) \leq max\_blue. \end{cases} \tag{5}$$

where $out\,(x, y)$ is the output pixel, $blue\,(x, y)$ and $green\,(x, y)$ is the value of the blue and green pixel respectively from the original image. The value of max_blue is the intensity value of blue component that are retrieved from histograms of original images, whereby it is calculated using the method shown in Fig. 2.

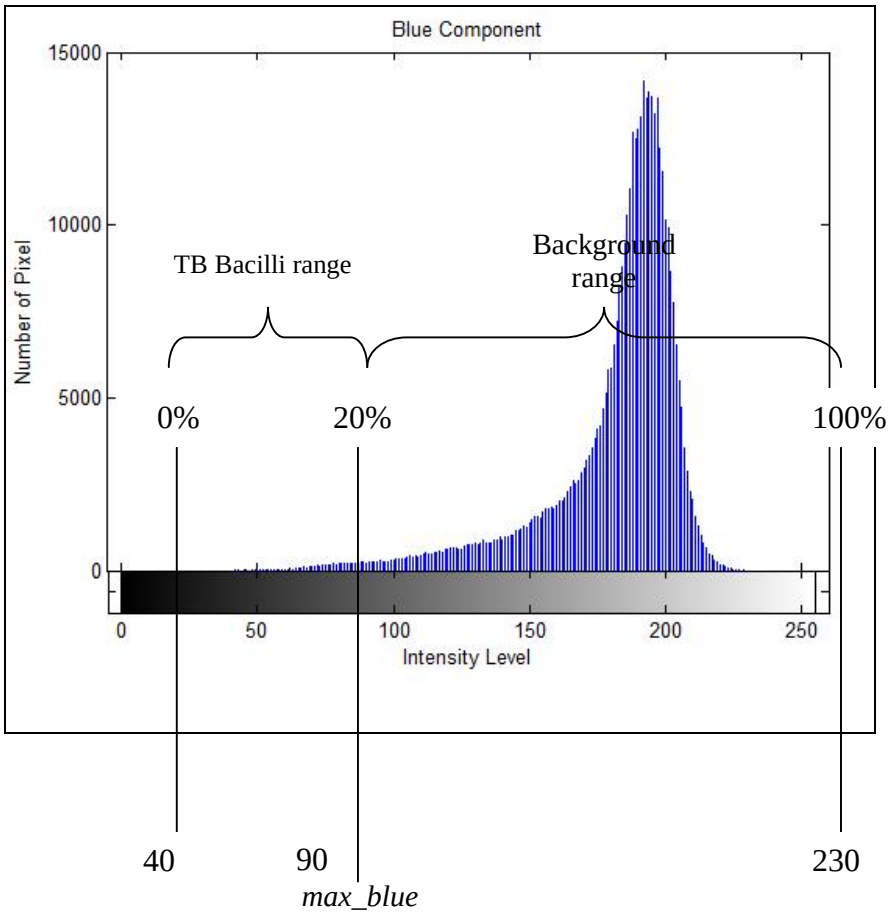


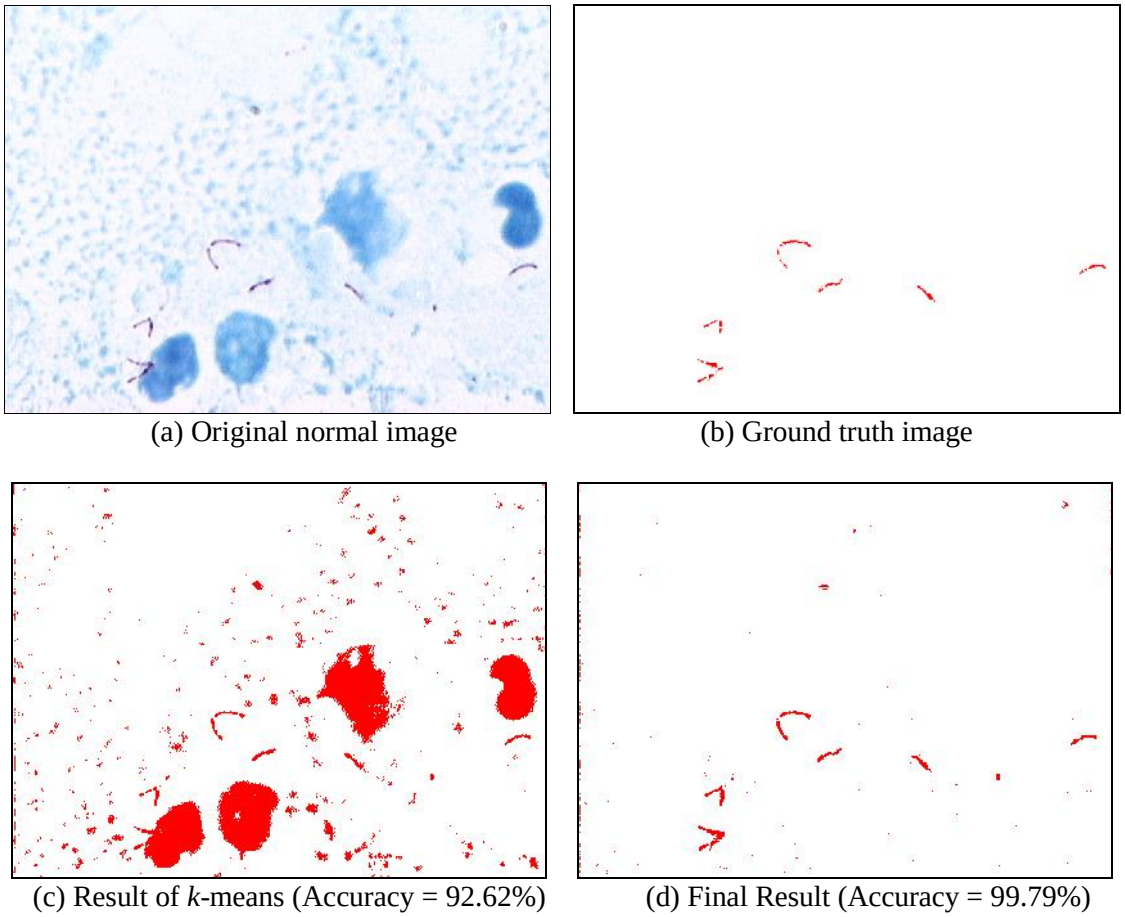
Figure 2. Method to retrieve *max_blue* value from blue component histogram

In Fig. 2, it can be seen that the sample of blue component histogram spread out in the intensity level range of approximately 40 to 230. Therefore, the point of 40 is marked as 0% and the point of 230 is marked as 100%. In this study, the value of *max_blue* is taken at the point where the histogram is marked as 20%. Therefore, from this histogram, the value for *max_blue* will be approximately 90. The point is marked as *max_blue* because it is the maximum value of blue component for that comprises of TB bacilli. A 20% value has been chosen from experimentation whereby it is found that from any ZN sputum slide images; only the lowest 20% of the blue component represents the TB bacilli pixels. However, this single rule cannot be applied on its own to segment the image. Therefore, it is used together with *k*-means clustering technique to optimize the segmentation accuracy.

The technique has been applied to the original raw images of Ziehl-Neelsen sputum slides. Testing images are taken from three various categories. For those images, set of clustering procedure was performed to segment the TB bacilli from the background. After the set of clustering procedure was applied, the pixels that belong to TB bacilli were highlighted in red colour, while those that belong to background were reduced to white.

3. RESULTS AND DISCUSSIONS

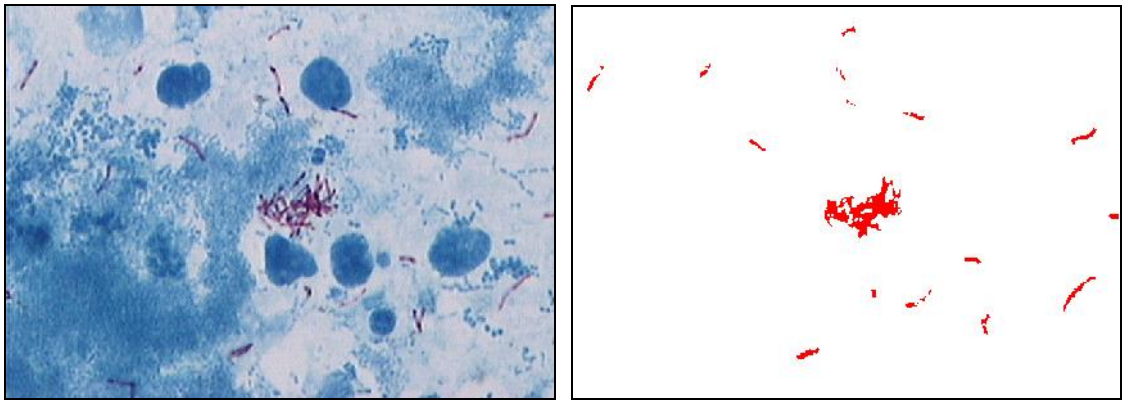
Fig. 3 - 5 show the segmentation results for several image of ZN sputum smear slide using *k*-means clustering. The original image is shown together with the ground truth image consisting of TB bacilli only and output image from applying the rule. The ground truth images are obtained by manually removing the background and sputum pixels using image editing software.



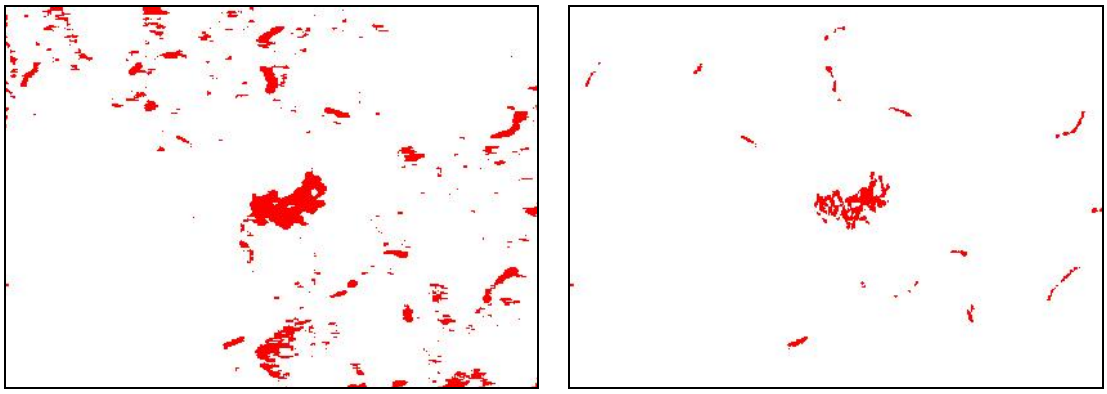
(a) Original normal image (b) Ground truth image

(c) Result of *k*-means (Accuracy = 92.62%) (d) Final Result (Accuracy = 99.79%)

Figure 3. Original and ground truth images from normal category with the result after applying set of clustering procedure

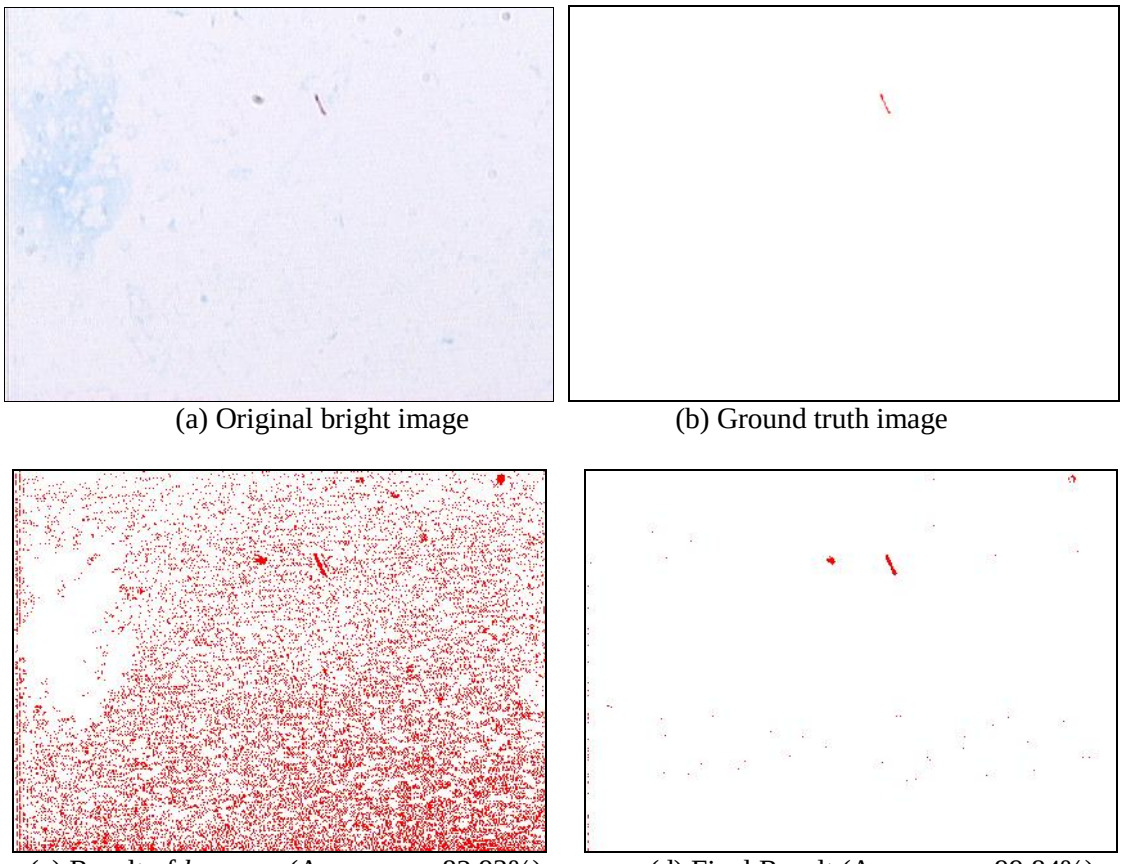


(a) Original dark image (b) Ground truth image



(c) Result of *k*-means (Accuracy = 96.15%) (d) Final Result (Accuracy = 99.38%)

Figure 4. Original and ground truth images from dark category with the result after applying set of clustering procedure



(a) Original bright image (b) Ground truth image

(c) Result of *k*-means (Accuracy = 82.93%) (d) Final Result (Accuracy = 99.84%)

Figure 5. Original and ground truth images from bright category with the result after applying set of clustering procedure

From the results presented, it can be seen that most of the background and sputum pixel have been eliminated after performing *k*-means clustering procedure. The results are considered in an acceptable accuracy whereby all images can reach accuracy of at least 99%.

4. CONCLUSION

In this section, a technique of image segmentation using original *k*-means clustering algorithm for Ziehl-Neelsen sputum slide images has been presented. The segmentation allows the elimination of a great amount of unwanted pixels, and retained only those pixels characterized to have similar colour to the TB bacilli. The values from the study conducted are assigned to be the initial centres for *k*-means clustering algorithm. Then *k*-means clustering have been applied to those images followed by several extra rules. The resulted images show encouraging results, which indicate that the proposed segmentation method is able to filter out the TB bacilli pixels from the background pixels.

REFERENCES

[1] WHO (2011). *WHO Report 2011: Global Tuberculosis Control*

[2] WHO (2007). *International statistical classification of diseases and related health problems, 10th revision (ICD-10)*, 2nd Edition.

[3] Salleh, Z., Mashor, M. Y., Mat Noor, N. R., Aniza, S., Abdul Rahim, N., Wahab, *et al*. (2007). Colour contrast enhancement based on bright and dark stretching for Ziehl-Neelsen slide images. *Proc. Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP 2007)*, 205-208.

[4] Osman, M. K., Mashor, M. Y., Saad, Z., & Jaafar, H. (2009). Contrast enhancement for Ziehl-Neelsen tissue slide images using linear stretching and histogram equalization technique. *Proc. IEEE Symposium on Industrial Electronics & Applications*, ISIEA 2009, 431-435

[5] Veropoulos, K., Campbell, C., & Learmonth, G. (1998). Image processing and neural computing used in the diagnosis of tuberculosis. *IEE Colloquium on Intelligent Methods in Healthcare and Medical Applications (Digest No. 1998/514)*, 8/1-8/4.

[6] Forero, M. G., Sroubek, F., & Cristobal, G. (2004). Identification of tuberculosis bacteria based on shape and color. *Real-Time Imaging, 10*, 251-262.

[7] Costa, M., Filho, F. C., Sena, J., Salem, J., & de Lima, M. (2008). Automatic identification of mycobacterium tuberculosis with conventional light microscopy. *Proc. 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2008*, 382-385

[8] Wang, W., Qin, Z., Rong, S., Rong, X., & Song, Y. (2008). A kind of method for selection of optimum threshold for segmentation of digital color plane image. *9th International Conference on Computer-Aided Industrial Design and Conceptual Design*, 959 – 961.

[9] Forero, M. G., Cristobal, G., & Borrego, J. A. (2003). Automatic identification techniques of tuberculosis bacteria. *SPIE Proceedings of the Applications of Digital Image Processing XXVI, 5203*, 71-81.

[10] Veropoulos, K., Learmonth, G., Campbell, C., Knight, B., & Simpson, J. (1999). Automated identification of tubercle bacilli in sputum: A preliminary investigation. *Analytical and Quantitative Cytology and Histology*, 21(4), 277–281

[11] Wilkinson, M. (1996). Rapid automatic segmentation of fluorescent and phase-contrast images of bacteria. *Fluorescence Microscopy And Fluorescent Probes*. New York, NY: Plenum Press.