Biomedical Engineering, Vol. 11, No. 1, 2014

AUTOMATIC SEGMENTATION OF MYCOBACTERIUM TUBERCULOSIS IN ZIEHL-NEELSEN SPUTUM SLIDE IMAGES USING SUPPORT VECTOR MACHINES

Selen Ayas, Murat Ekinci
Department of Computer Engineering, Karadeniz Technical University
Trabzon / Turkey
selenguven@ktu.edu.tr , ekinci@ktu.edu.tr

ABSTRACT

The World Health Organization suggests visual examination of stained sputum smear samples as a preliminary and basic diagnostic technique of tuberculosis disease. The visual examination requires laboratory technicians to spend considerable time, so it increases laboratorians' workload. In addition, it leads to a misdiagnosis because of requiring mental concentration. This paper presents a novel method for segmentation of tuberculosis bacteria in microscopic images taken from the Ziehl-Neelsen stained samples. Color information of bacterial regions which is taken from pixels and their adjacent pixels is sampled in training process. Multidimensional Gaussian probability density function and support vector machines are used during microscopic image segmentation comparatively. The performance of the implemented system is evaluated using sensitivity, specificity and accuracy criteria.

KEY WORDS

Tuberculosis, Ziehl-Neelsen staining procedure, Gaussian probability density function, support vector machines.

1. INTRODUCTION

Tuberculosis (TB) -one of the major health problems in the world- is an infectious disease caused by the bacillus *Mycobacterium Tuberculosis.* The bacillus seems like 1-10 micron lengthiness and 0.2-0.6 wideness, slightly curved or straight in microscopy. It has beaded and occasional branching form and also occur singly, pairs or in small clumps [1]. *Mycobacterium tuberculosis* and similar microorganisms have acid-fast cell wall which makes the cells impervious to acid-alcohol mixture. Therefore, acid-fast staining technique is used for detection of acid-fast bacilli (AFB). Ziehl-Neelsen (ZN) staining procedure is the most common method in acid-fast staining. AFB appears red-pink while non-acid-fast region is stained blue after staining with ZN procedure which is used by conventional microscopy [2]. Figure 1 shows an example of ZN-stained sputum smear image. Another staining procedure is fluorochrome staining in which bacillus is stained yellow fluorescence with dark background when observed with a fluorescence microscope [3]. Fluorochrome staining is more sensitive and require lower work effort than ZN staining. But the fluorescence microscopes are used in high-income countries because of greater cost of the equipment [4].

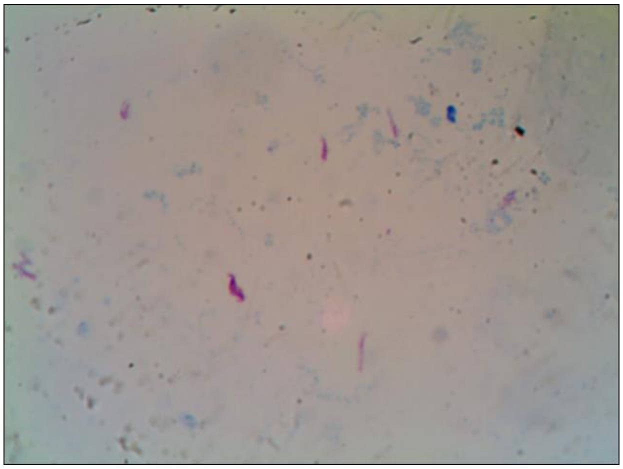


Figure 1. Example of ZN-stained sputum smear image

In TB suspected cases, patients' complaints, physical examinations, chest radiographs and tuberculin tests are not sufficient for a definitive diagnosis except for microbiology diagnostic [5]. In microbiology diagnostic, the tuberculosis is diagnosed by examining the stained sputum smear. The laboratory clinicians normally look for the presence of AFB in magnified microscopic images. Three specimens of sputum are drawn from the patient on two consecutive days and stained with ZN staining procedure. Experienced laboratory clinician needs to examine at least 100 field and spends at least five full minutes for each field. If each slide is not examined carefully or is examined too short, AFB will be missed and the specimen's result will be negative when it is actually positive [5]. Therefore manual screening is error-prone. Due to the process of visual examination requires mental concentration, number of specimens to be examined is limited. Additionally, it is a labour-intensive task because the examination of each specimen needs to visual examination for a long time [6]. On the other side, automatic screening speeds up diagnosis, reduces the workload of laboratory technicians and decreases error by improving accuracy and sensitivity of the diagnosis [7].

Several microscopic image segmentation techniques have been proposed for the identification of

DOI: 10.2316/J.2014.216.818-0055

mycobacterium tuberculosis bacillus. Osman M.K. *et al.* [8] used k-mean, moving k-mean and fuzzy c-mean clustering algorithms, and Otsu and iterative thresholding algorithms in C-Y color model. In yet another study, Forero *et al.* [9-11] proposed adaptive thresholding method in different color model. The method proposed in [12] performs to segment the bacilli by adaptive choice of the Hue range of HSI color model. As can be seen in the literature, most of these are related to basis thresholding or clustering methods and provide high segmentation accuracy. However, none of them apply to all microscopic images and different algorithms do not provide sufficient results for a specific image. Due to these reasons, to increase the accuracy of image segmentation, lots of intelligent approaches are suggested and it is benefited from knowledge of another fields such as artificial intelligence.

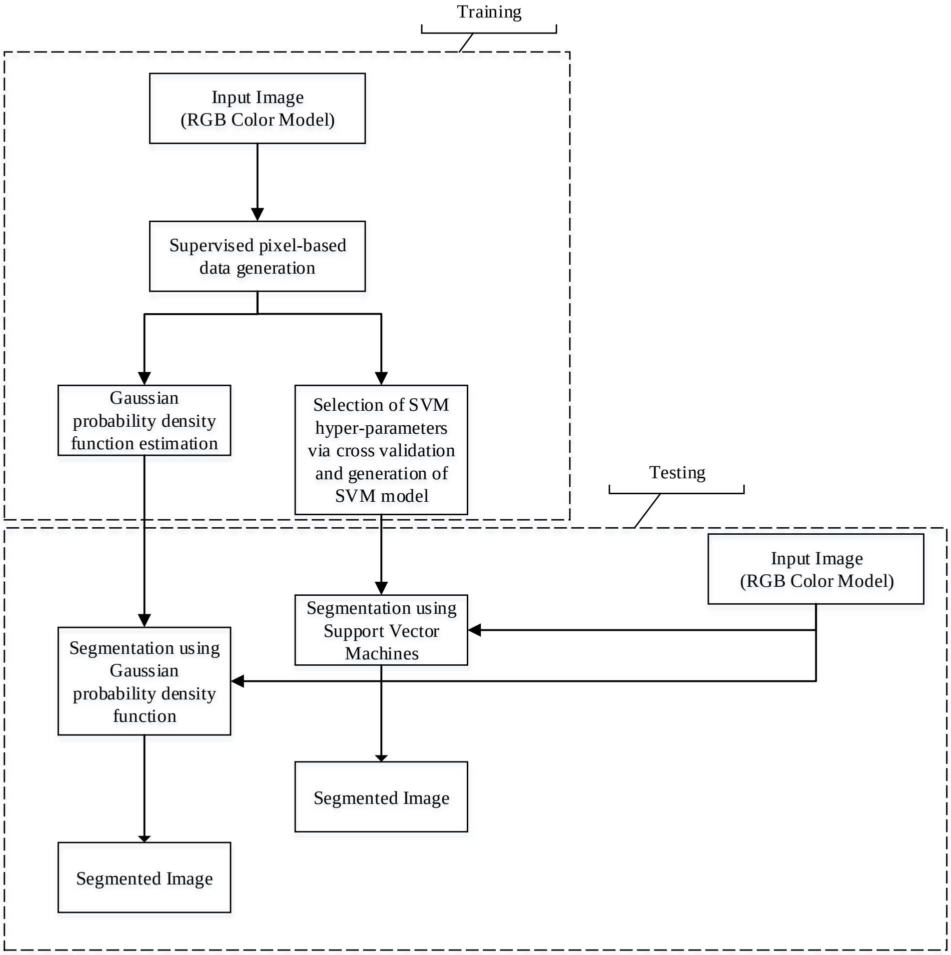

Training
Input Image
(RGB Color Model)
Supervised pixel-based data generation
Gaussian probability density function estimation
Selection of SVM hyper-parameters via cross validation and generation of SVM model
Testing
Input Image
(RGB Color Model)
Segmentation using Support Vector Machines
Segmentation using Gaussian probability density function
Segmented Image
Segmented Image

Figure 2. The flowchart of proposed method

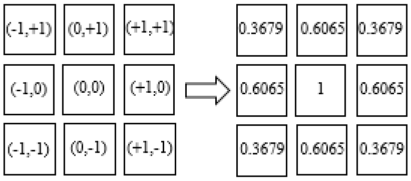
For these reasons, this study proposes an intelligent segmentation of tuberculosis bacilli in microscopic images. The ZN stained sputum smear slide images are analyzed and a pixel-based training approach is applied by a laboratory technician. The pixel-based segmentation involves the selection of bacterial regions pixel to pixel and segmentation of unanalyzed slide images by using Gaussian Probability Density Function and Support Vector Machines algorithms.

The flowchart indicating the segmentation process is shown in Figure 2. The rest of the paper is organized as follows: Section 2 gives information about methods used for feature extraction and classification. Section 3 explains our image acquisition system, database and, the experiments and the results of the proposed method. Finally, the paper concludes in Section 4.

2. THE PROPOSED METHOD

2.1 Feature Extraction based on Pixel Connectivity

In digital images, any pixel *p* has four horizontal and vertical adjacencies each of which has a unit distance from pixel *p*. It is named as four-connectivity and with the addition of four diagonal adjacencies, eight-connectivity adjacencies are obtained. In terms of pixel coordinates, each pixel with coordinates (x ± 1, y), (x, y ± 1) or (x ± 1, y ± 1) is connected to the center pixel located at (x, y) as shown in Figure 3(a).



(-1,+1)	(0,+1)	(+1,+1)
(-1,0)	(0,0)	(+1,0)
(-1,-1)	(0,-1)	(+1,-1)

0.3679	0.6065	0.3679
0.6065	1	0.6065
0.3679	0.6065	0.3679

Figure 3. (a) 8-connectivity adjacencies, (b) calculated coefficients

In the proposed approach, to extract the feature vector from bacterial pixels, it is taken advantage of pixels' color information and RGB color model is used. Therefore, red, green and blue component of the pixel in the position of the cursor and eight-connectivity adjacencies of this pixel are acquired. These components are redetermined using the coefficients which are obtained by fitting the bivariate GPDF to the center pixel and its eight-connectivity adjacencies. Bivariate GPDF for two uncorrelated random variables (x, y) is defined as follows;

$$f_x(x;y) = \frac{1}{2\pi\sigma_x\sigma_y} exp\left\{-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right\} \qquad (1)$$

where μ is mean and σ^2 is variance. Because of the fact that coefficient is constant, it's not regarded. If it is assumed that σ_x and σ_y are equal to 1, coefficients are calculated as in Figure 3(b).

2.2 Gaussian Probability Density Function

A random vector $X = [X_1 \ldots X_n]^T$ is said to multivariate normally distributed if its probability density function is defined as follows;

$$f_x(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\} \qquad (2)$$

where μ is mean vector, Σ is covariance matrix and n is the dimension of random vector. The mean vector is generated by averaging the each random variable X_i. It is the centroid of the probability density function or it is known as the point at which the probability density function is maximum.

In this work, the data which is extracted from training images in supervised pixel-based data generation step is represented by twenty-seven-dimensional Gaussian probability density function in our study. The range of a Gaussian curve is selected empirically by the user. The probability of belonging to bacterial region of the pixels in testing images is predicted by using this range and curve in segmentation process.

2.3 Support Vector Machine

Support Vector Machine (SVM) is a learning method which is used for classification and regression analysis. The basic idea behind it is to construct a maximum-margin hyperplane. So it means that SVM calculates the best hyperplane which separate the classes from each other. By using kernel functions, it maps pattern vectors to high dimensional feature space and separates data linearly in this space.

Decision function that uses the kernel function is defined as follows;

$$f(x) = sgn(\sum_{i=1}^{l} a_i y_i K(x, x_i) + b) \qquad (3)$$

where x is input vector, y is target value and $K(x,x_i)$ is the kernel function. The coefficients a_i and b are obtained from the following formula (4) which is required to maximise with respect to the a_i subject to (5).

$$max.\ L_D = \sum_{i=1}^{L} a_i - \frac{1}{2}\sum_{i,j=1}^{L} a_i a_j y_i y_j K\left(x_i x_j\right) \qquad (4)$$

$$subject\ to{:}\ 0 \le a_i \le C\ ,\ \forall_i = 1, \ldots L\ ,\ \text{and} \sum_{i=1}^{L} a_i y_i = 0 \ (5)$$

where C > 0 expresses the strength of penalty errors.

This decision machine method is applied to the training data acquired from microscopic images as follows:

1. A simple scaling is performed on the training data because of eliminating the computational complexity and transforming large numerical data into small numerical data.
2. Radial basis function is chosen as the kernel function. Because, this function maps the data to the high dimensional space and deals with the conditions in which the relation between features and labels is nonlinear.
3. To determine optimum C and γ hyper parameter, k-fold cross validation technique is used.
4. The training data are trained by using parameters determined in step (3).

3. EXPERIMENTS AND RESULTS

In this work, the ZN-stained sputum smear slides were prepared by Mycobacteriology Laboratory at Faculty of Medicine in Karadeniz Technical University. A smear-positive slide from a subject was used and 30 color images were acquired from it. Image acquisition system which is shown in Figure 4 was set up in Computer Vision and Pattern Recognition Laboratory [13].

Figure 4. Image acquisition system

This system consists of a standard PC, a microscopy and a digital camera. Sample slide was scanned by using a conventional light microscopy (Nikon Eclipse 80i) at 100x magnification. A Preimere Digital Microscope Eyepiece MA88-300 digital camera was attached to the ocular on a microscope for image acquisition. The taken images were stored in bitmap file format with 24 bit depth in color and the pixel resolution of an image was 640x480.

The whole data set consists of 30 positive images. To develop segmentation process, one third of these positive images were used for training and the rest of the images were employed to test the proposed technique. All images were analysed by a laboratory technician. So, it was decided whether each of the images' pixels belongs the bacterial region. One of these expert guided segmented images shown in Figure 5 were used for performance evaluation.

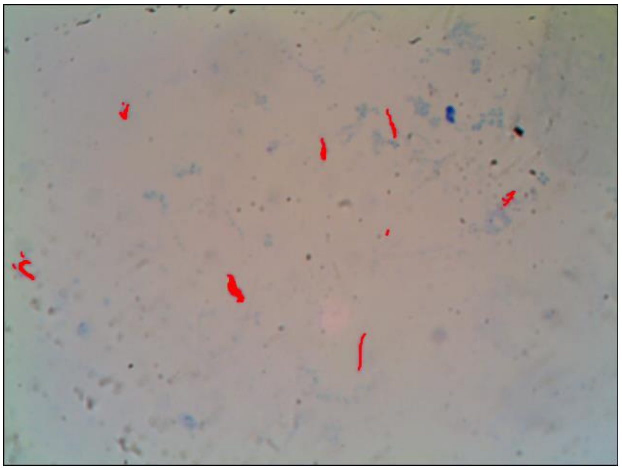


Figure 5. The expert guided segmented result image of Figure 1

In the proposed technique, the algorithm selected bacterial region pixel to pixel and 3x3 Gaussian mask was overlaid at the pixel location that corresponds to the mask's center. Then, nine dimensional feature vector for each color band (R, G, B) was calculated and so twenty-seven dimensional feature vector was obtained. To create feature vector for nonbacterial region, random numbers indicating the row and column indexes were generated. Color informations were extracted in this pixel coordinates as explained for bacterial regions. These feature vectors were trained and test images were classified with Gaussian probability density function and support vector machines classifiers.

The performance were estimated by using some criteria such as sensitivity, specificity and accuracy. For this reason the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) were obtained for each classifier. TP is the number of pixels which belong to bacterial region as they are, FP is the number of pixels which is not belonging to bacterial region but identified as belong to it, TN is the number of pixels which belong to nonbacterial region as they are and finally FN is the number of pixels which belong to bacterial region but identified as not belonging to it. Sensitivity measures the proportion of actual positive pixels which are correctly identified, specificity measures the proportion of actual negative pixels which are correctly identified and accuracy is the proportion of the number of correctly classified pixels to the number of pixels. This measures are given as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

$$Specificity = \frac{TN}{TN + FP} \tag{7}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

Table 1
Segmentation performance of Gaussian probability density function with different threshold values

Performance Measure	Threshold Value	%99.9	%99.8	%99.7
Sensitivity		%24.04	%55.94	%76.83
Specificity		%99.91	%98.61	%72.09
Accuracy		%99.62	%98.44	%72.08

Table 2
Segmentation performance of Support Vector Machines with optimum C and γ parameter and cross validation accuracy

Parameters		Performance Measure	
Optimum C	2048	Sensitivity	%87.02
Optimum γ	0.02352	Specificity	%99.59
Cross-validation accuracy	%99.46	Accuracy	%99.55

Table 1 and Table 2 shows the results of segmentation performance of these classifiers. As seen in Table 1, the range of Gaussian curve was selected empirically as %99.9, %99.8 and %99.7. While the threshold value is decreased, sensitivity rate increases. The reason why different threshold values weren't chosen after the value of %99.7 is that specificity decreases in value by approximately 26 percent. Because of this sharp decrease, the optimum threshold value was determined as %99.8 and the optimum sensitivity rate was %55.94. The segmented result image of Figure 1 by using Gaussian probability density function classifier is shown in Figure 6.

To apply support vector machine classifier, four steps represented in Section 2.3 were implemented firstly. Scaling values were chosen -1 as minimum and +1 as maximum. k parameter was chosen as 5. To determine optimum C and γ parameters, grid search approximation was used and, various pairs of (C, γ) were tried and the one with the best cross-validation accuracy was picked. In this approximation, C and γ parameters were tried in exponentially growing sequences (C= 2^{15}, 2^{13},…, 2^{-5} and γ= 2^{-6}, $2^{-5.5}$,…,2^{-1}).

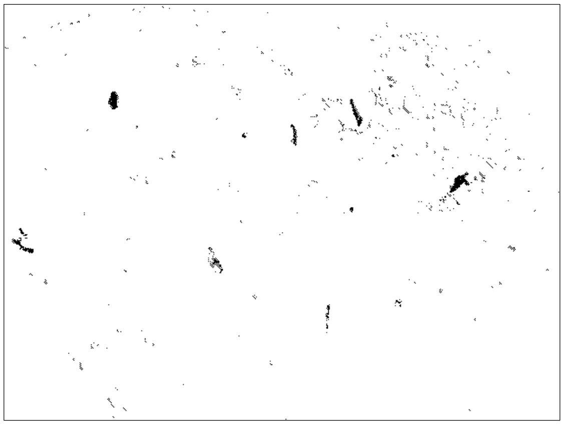


Figure 6. The segmented result image of Figure 1 by using Gaussian probability density function

As seen in Table 2, the best cross-validation accuracy was calculated as %99.46 when the optimum C and γ parameters were equal to 2048 and 0.02352, respectively. By using these parameters, sensitivity and specificity rates were calculated as %87.02 and %99.59, respectively. When these values were compared with the values calculated in Gaussian probability density funciton, it is shown that much better results were obtained in this classifier. The segmentd result image of Figure 1 by using support vector machines classifer is shown in Figure 7. When analyzing the segmented result images, the effect of the great difference between sensitivity rates is seen obviously.

The amount of time spent by the proposed algorithms are 0,021-0,592 sec and 0,020-24 sec for training and testing an image in GPDF and SVM, respectively. When this amount compared with the laboratorian's workload, it is clearly seen that the diagnosis time is reduced.

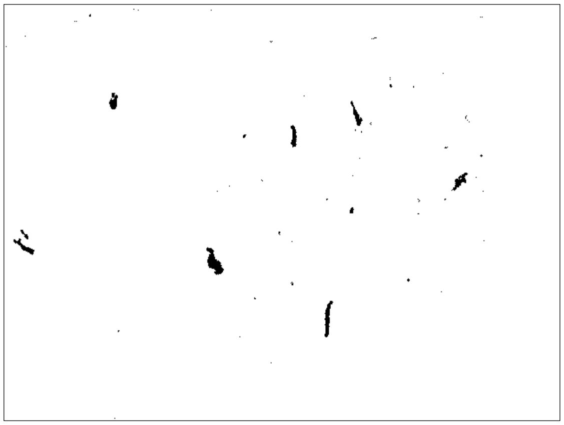


Figure 7. The segmented result image of Figure 1 by using support vector machines

4. CONCLUSION

A segmentation method has been proposed for tuberculosis bacilli in microscopic images of Ziehl-Neelsen stained sputum smears. Both of the quantitative and visual results of the classifiers are presented. It is obvious that much better performance is acquired by using support vector machines classifier. A comparison between the proposed algorithm and the other studies given in literature cannot be made because of the fact that the databases contain different images. Also, a shared database involves Ziehl-Neelsen sputum slide images is not available online. In the future work, different intelligent approaches may be performed and compared with the proposed approaches in this work and an online database structure will be constructed in order to make comparison for the other researchers.

REFERENCES

[1] J.C. Palomino, S.C. Leao & V. Ritacco, *Tuberculosis 2007-From basic science to patient care* (Bernd Sebastian Kamps and Patricia Bourcillier, 2007).

[2] International Union Against Tuberculosis and Lung Disease, *Sputum examination for tuberculosis by direct microscopy in low income countries* (2000).

[3] Auramine-rhodamine fluorescence -acid fast bacteria, http://www-medlib.med.utah.edu/WebPath/webpath.html.

[4] K. Steingart, M. Hnery, V. Ng, P. Hopewell, A. Ramsay, J. Cunningham, R. Urbanczik, M. Perkins, M. Aziz & M. Pai, Fluorescence versus conventional sputum smear microscopy for tuberculosis: a systematic review, *Lancet Infect. Dis., 6*, 2006, 570-581.

[5] Revised National Tuberculosis Control Programme, *Module for Laboratory Technicians* (Central TB Division, 2005).

[6] T. N. L. Nguyen, C. D. Wells, N. J. Binkin, D. L. Pham & V. C. Nguyen, The importance of quality control of sputum smear microscopy: the effect of reading errors on treatment decisions and outcomes, *Int J Tuberc Lung Dis, 3*(6), 1999, 483-487.

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

[8] M. K. Osman, M. Y. Mashor & H. Jaafar, Performance comparison of clustering and thresholding algorithms for tuberculosis bacilli segmentation, *2012 International Conference on Computer, Information and Telecommunication Systems*, Amman, 2012, 1-5.

[9] M. G. Forero, G. Cristobal & J. Alvarez-Borrego, Automatic identification techniques of tuberculosis bacteria, *Applications of digital image processing XXVI, 5203*, 2003, 71-81.

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

[11] M. G. Forero, G. Cristobal & M. Desco, Automatic identification of mycobacterium tuberculosis by Gaussian mixture models, *Journal of Microscopy, 223*(2), 2006, 120-132.

[12] V. Makkapati, R. Agrawal & R. Acharya, *IEEE International Conference on Automation Science and Engineering*, Bangalore, India, 2009, 217-220.

[13] http://ceng2.ktu.edu.tr/~cvpr