Automatic identification techniques of tuberculosis bacteria

Manuel G. Forero\textsuperscript{a}, Gabriel Cristóbal\textsuperscript{a} and Josué Alvarez-Borrego\textsuperscript{b}

\textsuperscript{a}Instituto de Óptica (CSIC), Serrano 121 28006, Madrid, Spain
\textsuperscript{b}CICESE, División de Física Aplicada, Dept. de Optica, Carretera Tijuana-Ensenada, B.C. México

ABSTRACT

Tuberculosis is a serious illness which control is based on early diagnosis. A technique commonly used consist of analyzing sputum images for detecting bacilli. However, the analysis of sputum is quite expensive, time consuming and requires highly trained personnel to avoid high errors. Image processing techniques provide a good tool for improving the manual screening of samples. In this paper we present a new bacilli detection technique with the aim to attain a high specificity rate and therefore for reducing the time required to analyze such sputum samples. This technique is based on the heuristic knowledge extracted from the bacilli shape contour. It uses also the color information for image segmentation and finally a classification tree is used to categorize if a sample is positive or negative.

Keywords: color, fluorescence microscopy, pattern recognition, tuberculosis.

1. INTRODUCTION

Mycobacterium tuberculosis bacilli is the origin of the pulmonary tuberculosis disease, although these organisms can infect other organs or tissues such as the brain, kidneys, bone and skin. Tuberculosis is the main cause of death produced by an infectious illness. According to the World Health Organization a third part of the worldwide population (1722 million people) are carrier of this bacterium, originating 10 million cases of active tuberculosis worldwide and approximately 3 million of deaths annually\textsuperscript{1}. Some studies conducted in Spain shown that cigarette smoking is a risk for pulmonary TB in young people\textsuperscript{2}. An epidemiology study of the tuberculosis in the USA present a Markovian model nature showing an increase in the ensuing decade and a decline again\textsuperscript{3}. Some of the reasons for the increase incidence of this disease are the HIV infection (tuberculosis is the main cause of death in HIV infected people) and multidrug resistance.

Besides of clinical suspicion, the diagnosis of mycobacteriosis must be done through genus specific smears of clinical specimens. Identification of tubercle bacilli are routinely done in sputum smears using a fluorescence microscope dyed with fluoroscope auramine. These techniques lack of sensitivity and consequently clinicians must wait culture results as much as two months because this bacilli takes 5 to 20 hours to duplicate itself. Manual screening for the bacillus identification involves a labor intensive task with a high false negative rate\textsuperscript{4}. Automatic screening will entail several advantages, like a substantial reduction in the labor workload of clinicians, improving the sensitivity of the test and a better accuracy in diagnosis by increasing the number of images that can be analyzed by the computer.

Bacteria segmentation of particular species entails a complex process. Bacilli shape is not enough as a discriminant feature, because others bacteria, species and particles share the same morphology. Therefore besides shape the color bacilli information was used.

Several works have been addressed to the segmentation of bacteria particles. Veropoulos et al\textsuperscript{4, 5} used an identification method based on shape descriptors and neural network classifiers showing a sensitivity (ratio of true positive decisions against the total number of positives cases) of 94.1\%. Wilkinson\textsuperscript{6} proposed a rapid multiresolution segmentation technique based on computing different thresholds for different areas of a grey level image. Other authors use the color information as the key discriminant factor either for bacteria segmentation and identification\textsuperscript{7, 8} or cell segmentation in lung cancer diagnosis\textsuperscript{9, 10}.

As a follow-up of previous works\textsuperscript{11, 12}, in this paper a new technique for bacilli segmentation and identification is described in order to establish if an image is positive or negative. As above mentioned, the segmentation method is based on the use of the color information and the identification technique requires the use of several descriptors that have been calibrated using the heuristic knowledge about the most frequent bacilli shapes.
Staining procedures were performed with both respiratory and non-respiratory clinical specimens except for urine and blood ones. These specimens were stained with the fluorochrome auramine O and were scanned by using a fluorescence microscope at x250 magnification. Because acid-fast artifacts may be present in a smear, it is necessary to review the cell morphology carefully. The confirmation of positive fluorochrome smears was made with the growth of M. tuberculosis bacilli from the culture of specimens in liquid and solid media. Sample slides were analyzed with a Zeiss Axiophot photomicroscope illuminated with a Zeiss Attoarc variable intensity fluorescence illumination system. For image acquisition a Coolsnap digital camera from Photometrics was used.

A total of 397 negative and 75 positive images were acquired. To develop the identification process, 110 bacilli extracted from 15 positive images were analyzed and the other 60 were employed to verify the proposed technique.

3. METHOD

If the M. tuberculosis bacilli appear in an image, they fluoresce in the range between green and yellow up to white. The bacilli measure approximately 1 to 10 µm in length and 0.2 to 0.6 µm in width and they can present a straight, curved or bent shape. Individual bacilli may display heavily stained areas and zones of alternated stain producing a beaded appearance. This information is important in the segmentation and identification processes as it will see later. Figure 1 summarizes the learning method. It is important to emphasize that we are more interested in a correct test evaluation procedure and not in the correct identification of all bacilli.

The first stage of the process is a segmentation process. The segmentation procedure consists of several steps. First, similarly to the method proposed by,\textsuperscript{4,5} a Canny operator (\( \alpha = 1 \)) is used to detect the borders of the image (see Fig. 2.b). This step is followed by a non-maxima suppression and a hysteresis threshold operations.

Because some structures can appear broken, a morphological closing operation is applied. As it can be observed in Fig. 2.c, several images still have some open edges non belonging to bacteria. These edges are normally one pixel wide, therefore in order to eliminate them all the close contours are filled out and then a morphological opening is applied.

Due to the fact that the images are in RGB color format, it is necessary to study their spectral behavior. After scrutinizing these images, the following facts can be observed:\textsuperscript{11}:

- Both the regions corresponding to the image background and bacilli appear with similar color, intensity and texture.
- The bacilli show fluoresced green to yellow color and in some cases it could be white.

Also after observing the color channel histograms and the bacilli color profile one can conclude that the bacteria appear dyed with green color showing a high intensity on the green channel and very low intensity in the others. The yellowed bacilli show a high and similar intensity in the green and red channels. Finally, when the bacilli appear white, the intensity in all the three channels are similar and present a high value. It
was observed too, that in all cases there is at least a pixel where the green value is bigger that the red one. In addition, it was found that the blue channel does not entails significant information for distinguishing the bacilli, what is in accordance with the results of Demantova et al.\textsuperscript{8}

Based in these facts and in the statistical analysis of the sample’s intensities, a color segmentation technique was experimentally established. An object will be retained if it has at least a pixel whose green value is higher.
than 180 and bigger than the corresponding value in red channel. Otherwise it is rejected. Figures 2 to 4 show the results obtained with this technique applied to two negative images and a positive image.

Once the image has been segmented, an identification process will be performed in order to determine if the remaining objects are bacilli, and therefore if the image is positive or negative. Several methods can be employed to classify objects, including neural networks (proposed i.e. by Veropoulos\textsuperscript{4,5}) and clustering techniques.
Figure 4. Positive sample. 

Before to apply the identification process, the remaining objects are filtered according with the number of pixels of their contour, rejecting those objects whose contour is too small or too big to be bacilli.

Later, different descriptors were evaluated for the bacilli characterization: area, compactness, major axis length, minor axis length, eccentricity, perimeter, solidity, the 7 Hu’s moments and the first 16 normalized Fourier descriptors. The representation of the bacilli must be preferably invariant against translations, rotations and scale changes in order to identify the tuberculosis bacilli. Then, a data set of reference bacilli were analyzed, with the purpose of studying and drawing conclusions about it. This process is developed to establish the distribution of the samples, determinant to define the later steps in the classification, besides of establishing whether the number of taken samples is adequate and representative for diminishing the error rate (see Fig. 5).

By means of a previous visual analysis it was observed that the bacilli do not have an uniform shape, and that they can be subdivided in clusters according to their shape, thickness and length. In addition, several objects in spite of being bacilli, have an ovoid shape that, in general, is not characteristic of the M. tuberculosis and therefore they can be confused with another object, as the cases illustrated in Fig 6. These facts are in accordance with the information provided by a microbiologist expert, who indicates that although some bacilli are not easy to distinguish by their shape and color when they appear isolated, they are classified as bacilli if at the same image other bacilli can be easily recognized. Another fact that can increase the classification error are due to debris objects that often appear in zones near the decision regions. In this way, the amount of objects was reduced to 88 of the 110 described in Section 2. It can be observed that the rejected objects were well characterized by their low compactness and eccentricity.

![Figure 5](image_url). Data set of reference bacilli used in the present study

![Figure 6](image_url). Samples of rejected bacilli due to their shape.

In order to determine the number of clusters that better define the samples a clustering algorithm was employed. Seven clustering algorithms were tested: adaptive, chained distances (chain map), K-means, sequential,
max-distance also known as max-min and Batchelor and Wilkins, ISODATA and matrix of similarity. Each algorithm is controlled by different parameters and heuristics, which provides a great flexibility to whole process.

Simultaneously and in order to retain and recognize the most useful descriptors an analysis and screening process was conducted. Because not all the descriptors have good discrimination ability, their were selected according with their mean and variance. Assuming that the descriptors have a Gaussian distribution, the distributions were displayed for different number of classes and it was observed the presence of overlapping between the different classes. As it can be observed in Figure 7, the first 4 Hu’s moments, the eccentricity and the compactness provide a good separation between the bacilli clusters. For this reason they were selected to make the classification.

In addition, in the K-means clustering technique, in which the only parameter defined by the user is the number of clusters to be located, a silhouette method was applied and the coefficient of silhouette calculated for the resulting clusters for a rank between 3 and 10. To avoid the bias due to the election of the initial centroids that usually can be observed, the centroids were calculated several times. This was performed by selecting different objects as initial centroids and then comparing the results obtained until a minimum is found. As result of these analysis 4 classes were found. Figure 8 shows the silhouette of the four classes, in which the silhouette coefficient is 0.702. This value indicates that a reasonable structure has been found. The silhouette value of an object measures its similarity with respect to the other objects in its own cluster, compared to objects located in other clusters.

Once the classes were found, a first classification was made by using the nearest neighbor method. The best results were obtained with the 4 first Hu’s moments and the compactness. The Mahalanobis distance was used as similarity criterion due to the fact the magnitude of the descriptors and their variance are not similar.

Given the nature of the objects to identify, we have a heuristic description of the shape of bacilli that can be also used to make the classification. This information is related with the width and length of the bacilli. Preliminary tests limited to the use of the obtained centroids did not allow to obtain good results, since several objects different from the bacilli, present descriptors very similar to the bacilli and therefore they are classified as bacilli. The compactness and the eccentricity are two complementary descriptors that allows to know whether an object is long and thin which is a typical shape of the bacillus as mentioned before. Using these values and the Hu’s descriptors, the classification tree was constructed[13] (see Fig. 9).

Besides the previous classification tree, some additional nodes can be defined with the purpose of increasing the sensibility. The reason for such new nodes is due that eventually it can appear bacilli whose length is greater to the expected one. Therefore the system is not able to classify them as bacillus, since its shape is not near anyone of the 4 found clusters. However from the limited number of positive samples it was difficult to determine the exact compactness and eccentricity of them. Therefore, some of these bacilli shapes were simulated with the purpose of complete the present study. This point requires to be confirmed with the availability of positive samples. Figure 10 shows the classification tree 2, that takes into account the presence of these longer bacilli.

4. RESULTS

Figures 2-4 show that the segmentation technique allow to extract most of the bacilli in an image, eliminating most of the rest of objects that are not bacilli. In some negative images the segmentation identification process eliminates all objects and therefore the remaining identification process is unnecessary for assigning the image as negative. Table 1 shows results of the specificity and sensibility of the current method, including the confidence intervals.

<table>
<thead>
<tr>
<th>Images</th>
<th>Specificity</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree 1</td>
<td>93.70 ± 2.39%</td>
<td>93.3 ± 6.31%</td>
</tr>
<tr>
<td>Tree 2</td>
<td>91.68 ± 2.71%</td>
<td>100%</td>
</tr>
</tbody>
</table>
Figure 8. Silhouette of the 4 obtained clusters.

Figure 9. Classification tree 1. Each left branch is taken if the node is true and the right one if it is false. B means bacillus and R a rejected object. d is the Mahalanobis distance between the object and the nearest cluster, c is the compactness, e the eccentricity, t is a distance threshold that was fixed to 10.

From the results shown in the table, it can be seen that the developed technique appears as a viable solution for the identification of bacilli in sputum samples. This method provides better sensitivity rate for a similar specificity in comparison with the results reported by Veropoulos. As a future research, we can mention that in order to verify the feasibility of the current method in the clinical routine, it is still necessary to check the specificity for each set of images corresponding to each microscopic sample (between 10-100 images per sample) and not in terms of individual images.

5. CONCLUSIONS

A new technique for analyzing fluorescence images of sputum was presented. The technique is based in new method for segmentation followed by an identification procedure. The segmentation allows the elimination of a great amount of debris objects, and therefore only those characterized to have by a similar color than the bacilli are retained. This screening process provides a significant reduction in the computational complexity reducing the search space of candidate particles to those which color fits with the expecter Mycobacterium tuberculosis
color. One of the key stages of the current system is analysis and screening of the bacilli shapes. In this work 110 samples of bacilli were analyzed. It can be expected that with the availability of a more ample dataset the performance of the system can be improved and also the confidence interval can be reduced. Although the descriptors used to calculate the cluster centroids are appropriated for a recognition system, however in some cases the classification failed. Therefore, we used the heuristic knowledge about bacilli to construct a classification tree and in this way the final results were improved. On the other hand, the heuristic knowledge used to build up the classification tree was tested to construct a rule-based fuzzy logic system. Nevertheless, the results attained so far with the fuzzy system have worst performance than the method described here.

6. ACKNOWLEDGEMENTS

This work has been partially supported by the following grants: TIC2001-3697-C03-02; III Pricit of the Comunidad Autonoma de Madrid and the IM3 medical imaging thematic network. M.G. Forero-Vargas is supported by a Spanish States Secretary of Education and Universities fellowship.

REFERENCES


