A FEATURE DETECTION METHOD FOR PULMONARY MYCOBACTERIUM TUBERCULOSIS USING CONVENTIONAL CHEST RADIOGRAPHS

Norliza Mohd. Noor, Omar Mohd. Rijal**, & Chang Yun Fah****

** Dept. of Electrical Engineering, Universiti Teknologi Malaysia City Campus*
***Institute of Mathematical Science, Faculty of Science, University Malaya*
****Faculty of Engineering, Multimedia University*

ABSTRACT

The success of eliminating the disease Mycobacterium Tuberculosis (MTB) depends on the detection capabilities of medical organizations. In Malaysia, the government hospitals perform the major part of this particular task. An important ingredient of the diagnostic process in government hospital is the visual interpretation of standard chest X-ray films. A previous study proposed an objective alternative; involving wavelets coefficient, as the feature vector of MTB. In this study, we proposed an Andrew's Curve graphical presentation of the feature vector of MTB.

INTRODUCTION

Due to economic considerations the conventional x-ray film is still an important ingredient in the diagnostic process despite rapid advances in medical imaging technology (see for e.g. Middlemiss) [1] and Moores [2]. A description on the reliability of chest radiography is discussed in Toman [3].

The authors in [4] looked at subsets of the digitized chest x-ray image of a confirmed MTB patient. In particular, the infected region seen visually as white cloudy or snowflakes is the region of interest in this study. We then take some sample of grey-level values or pixel values the infected areas. The samples are in the form of vertical lines defined between a given pair of adjacent ribs, which in turn is defined as the line profile.

Thirty line profiles were obtained. For each line profile, applying one-dimensional discrete wavelet [5] gave the corresponding approximate and detailed Daubechies Coefficients. In total, a vector of 26 coefficients represented each line profile. Hierarchical clustering techniques [6] were applied using Minitab [7] and SPSS [8].

THE PILOT STUDY

In the pilot study [4] we studied two sets of data, confirmed tuberculosis patients and none-MTB patients. The chest radiographs of the confirmed MTB patients were provided by the Respiratory Unit, Kuala Lumpur Hospital and none-MTB patients were provided by Selayang Hospital.

The medical expert on MTB identified the infected area. For each infected area, we sampled 30 lines profile. For each line profile, we obtained 26 Daubechies coefficients. Six hierarchical clustering techniques were applied to the 30 x 26 approximate Daubechies coefficients. The general result of clustering showed that most techniques separate the line profiles into two groups: the first set nearest to the 'wind-pipe' is regarded as primary infected area, whilst the other set may be considered the secondary infected area.

Figure 1(a) shows a chest X-ray of the confirmed MTB patients. Figure 1(b) shows the line profiles selected. As an example of clustering, Figure 2a(i) is the dendogram using complete linkage and Figure 2a(ii) a schematic representation of line profiles being separated into two groups. The grouping of line profiles is summarized in Table 1. We define the average of the profile vectors, for example for the complete linkage;

$$\bar{\underline{x}} = \frac{\underline{x}_1 + \underline{x}_2 + \underline{x}_3 + \underline{x}_{11} + \underline{x}_{12} + \underline{x}_{22} + \underline{x}_{23} + \underline{x}_{24} + \underline{x}_{25}}{9} \quad (1)$$

where $\underline{x}_j$ is the vector of approximate coefficients for the j^{th} line profile.

From Table 1, we could see that clustering using the complete linkage, Between Group Average and Within Group Average gave the same grouping of line profiles. We denote the average for these vectors as X_A. Also, X_B, X_C and X_D represent the average vector for Ward, Centroid and Median method,

respectively. The Euclidean's distance between these vectors are shown in Table 2. Table 2 indicates the $\bar{\underline{x}}$ -vector for all clustering method is similar. Henceforth the $\bar{\underline{x}}$ -values, (for example complete linkage) may be used as a feature to identify MTB.

In summary, clustering of line profiles, or equivalently clustering of vectors of approximate Daubechies wavelet coefficients may be used as a method to identify regions that are infected with MTB. The average value of the profile vectors, for example, for complete linkage is as follows:

$\bar{\underline{x}}$ = (2.456, -14.723, 51.464, 151.881, 149.618, 148.203, 146.751, 148.195, 145.076, 144. 298, 145.446, 144.347, 146.718, 145.889, 146.373, 149.406, 146.237, 145.709, 147.076, 147.818, 150.784, 150.092, 149.977, 143.548, 170.145, 24.881)

Therefore $\bar{\underline{x}}$ may be used as a feature characteristic of the MTB disease.

ANDREW'S PLOT

Whilst the $\bar{\underline{x}}$ -vectors given in equation (1) may be used as a feature to identify MTB, there is a need to be able to compare the $\bar{\underline{x}}$ -vectors, for example:

(i) To compare two X-ray films of a patient undergoing treatment after one month.

(ii) Comparing a 'new' patient with a confirmed MTB patient.

The visual comparison of two 26-dimensional vectors is clearly not appealing. Andrews [9] proposed a method of plotting a data point $\underline{x}_r^T = (x_{r1}, \ldots, x_{rp}), r = 1, \ldots, n$, which involves plotting the curve $\{t, f_{\underline{x}_r}(t)\}$ where

$$f_{\underline{x}_r}(t) = x_{r1}/\sqrt{2} + x_{r2}\sin t + x_{r3}\cos t + x_{r4}\sin 2t + x_{r5}\cos 2t + \cdots \quad (2)$$

for each "data point' $\underline{x}_r (r = 1, \ldots, n)$ over the interval $-\pi < t < \pi$. Thus, each data point (in this case our $\bar{\underline{x}}$-vectors) will appear as a harmonic curve drawn in 2 dimensions. It may be shown that $\int_{-\pi}^{\pi} [f_{\underline{x}}(t) - f_{\underline{y}}(t)]^2 dt$ between two curves $\{t, f_{\underline{x}}(t)\}$ and $\{t, f_{\underline{y}}(t)\}$ is proportional to the square of Euclidean distance between $\underline{x}$ and $\underline{y}$.

CLUSTERING AND ANDREW'S PLOT

Since the $\bar{\underline{x}}$ -vector are very similar, we propose studying $\underline{v} = (\bar{\underline{x}}_1 + \bar{\underline{x}}_2 + \ldots + \bar{\underline{x}}_6)/6$ where the subscripts 1,2, …,6 represent the six clustering methods for a given patient. This is done in accordance with a standard approach in statistical clustering, namely accept the clusters suggested by the majority of clustering methods. In this study the Andrew's Curve of vector $\underline{v}$ were considered.

A random sample of ten patients were selected; six confirmed MTB patients from The Respiratory Unit, Kuala Lumpur General Hospital and four non-MTB patients from The Selayang Hospital.

The Andrew's Curve for each of the ten patients or data points $\underline{v}_r (r = 1,2,\ldots,10)$ were compared for values of t between zero and six. These values of t were chosen solely for producing graphs that may be recognized and differentiated with ease.

One confirmed MTB patient (black curve) and one non MTB patient (blue curve) were randomly selected and compared in Figure 3. Both curves show similar trend, except that the amplitude of the black curve is larger than the blue curve.

To show that the shape of the curves in Figure 3 is not a chance occurrence, Figure 4 compares the same blue curve with all the six confirmed MTB patients. Clearly, the black curves are clustered in a group, distinct from the blue curve. Further, in Figure 5, one confirmed MTB patient is compared with the other four non MTB patient showing similar result.

DEVELOPMENT OF AN MTB FEATURE DETECTION SYSTEM

Both Conventional chest X-ray and digital chest x-ray were used in this study. The x-ray films were digitized using film scanner and transfer to a PC-based system. The MTB detection software was develop in MATLAB 6.1 GUI (Graphical User Interface) environment [10]. The image of digitized chest X-ray is first

displayed on the screen. Then the user will be prompt to select the area that need to be analyzed. The selected area will then be display on the screen, the user then will need to provide some data for testing by drawing with the cursor several lines at the suspected infected area. The software will then generate line profiles and calculate the Daubechies approximate and detail coefficient. The Daubechies approximate coefficients were used as feature vector in this research. All the feature vector of the lines profile was then subjected to 6 clustering methods: complex linkage, centroid method, median method, Ward's method, between group average and within group average.

The Dendogram (see Figure 2) and a schematic representation of the clusters are displayed. Finally the Andrew's plot may also be displayed for selected values of t and $f_v(t)$.

SUMMARY AND FURTHER REMARKS

A pilot study of an objective method for feature detection for MTB is proposed whereby the usual problems associated with visual interpretations of images are removed. Apart from detection, the system allows the medical practitioner to 'explore' the image and perform segmentation.

However, the robustness and the sensitivity of the method still need to be studied in the sense that a larger database (more patients) should be obtained and compared.

ACKNOWLEDGEMENT

We would like to acknowledge cooperation given by:
(i) Dr. I. Kuppusamy and Dr. Azwayati Abas from The Respiratory Unit, Kuala Lumpur General Hospital,and

(ii) Dr. Rosnah Hadis and Dr. Zaharah Musa from The Selayang Hospital.

This research was funded under short-term research grant from Research Management Centre, UTM.

REFERENCE

[1] Middlemiss, H, "Radiology of the future in developing countries", British Journal of Radiology, 55, pp. 698-699, 1982.

[2] Moores, B.M., Digital X-ray Imaging, IEE Proceedings, Vol. 134, part A, Number 2, Special Issues On Medical Imaging, 1987.

[3] Toman, K., Tuberculosis: Case Finding and Chemotherapy, W.H.O. Report, pp. 28, 1979.

[4] Noor, N. M., Rijal, O.M and Chang, Y.F., "Wavelet as features for Tuberculosis (MTB) using standard x-ray film images", IEEE proceedings of 6th International Conference on Signal Processing (ICSP02), Beijing, China, 2002.

[5] Daubechies, I., Ten Lectures on Wavelets Society For Industrial and Applied Mathematics, Philadelphia, 1992.

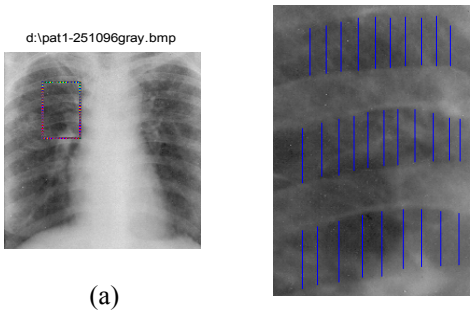
[6] Everitt, B. S., Cluster Analysis, Heinemann Educational Books Ltd, London, 1977.

[7] MINITAB Software for Windows, The MINITAB Inc.

[8] SPSS Software for Windows, The SPSS Inc.

[9] Andrews, D.F., "Plots of high dimensional data", Biometrics, 28, pp.125-36, 1972.

[10] Matlab software, The Language of Technical Computing, The Mathworks Inc.



(a) (b)

Figure 1: (a) A chest X-ray of a confirmed MTB patient; (b) A subset of (a) showing the line profiles taken between the area of 2nd, 3rd, 4th and 5th ribs.

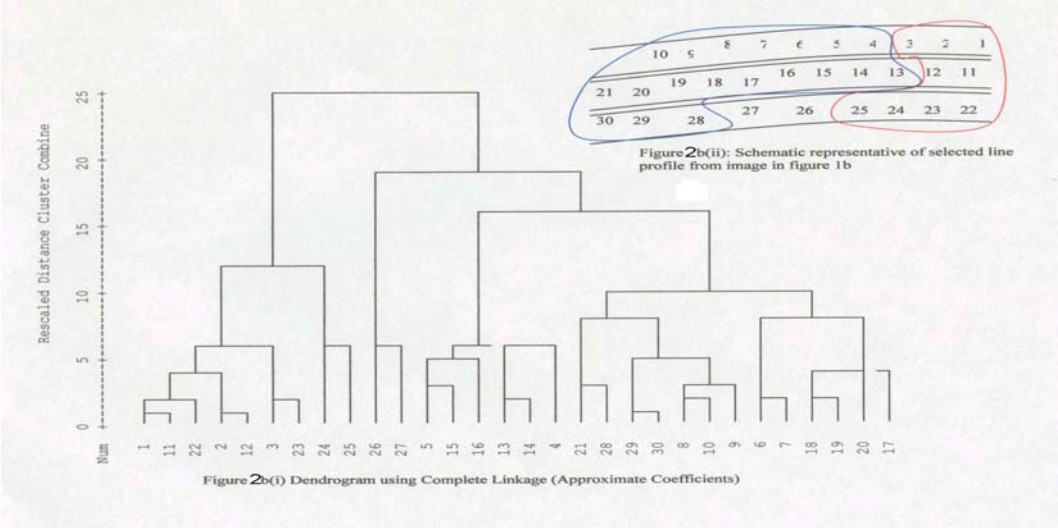


Figure 2b(ii): Schematic representative of selected line profile from image in figure 1b

Figure 2b(i) Dendrogram using Complete Linkage (Approximate Coefficients)

Figure 2

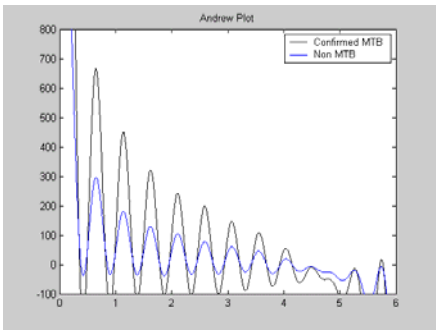


Figure 3

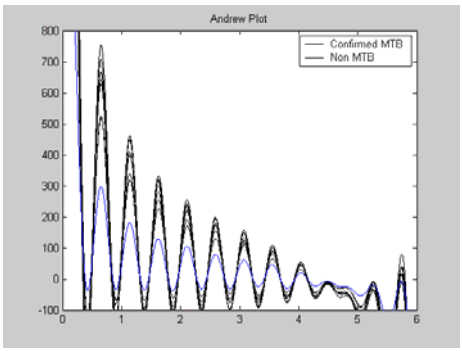


Figure 4

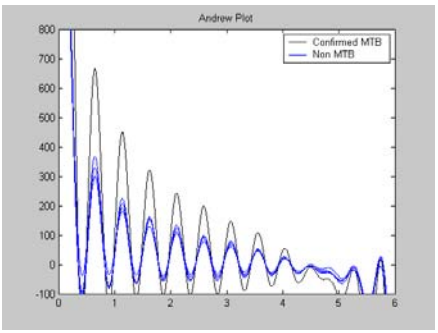


Figure 5

Table 1: Summarized grouping of line profiles using six hierarchical clustering method.

Clustering Method	**Grouping of line profile**
Complete Linkage	1, 2, 3, 11, 12, 22, 23, 24, 25
Centroid Method	1, 2, 3, 11, 12, 22, 23
Median Method	1, 2, 3, 4, 5, 6, 7, 11, 12, 13, 14, 15, 16, 22, 23
Ward's Method	1, 2, 3, 4, 11, 12, 22, 23, 24, 25
Between Group Average	1, 2, 3, 11, 12, 22, 23, 24, 25
Within Group Average	1, 2, 3, 11, 12, 22, 23, 24, 25

Table 2: Euclidean's Distance Matrix for X_A, X_B, X_C and X_D.

	$\mathbf{X}_A$	$\mathbf{X}_B$	$\mathbf{X}_C$	$\mathbf{X}_D$
$\boldsymbol{X_A}$	0	0.000484730	0.005712382	0.020888985
$\mathbf{X}_B$	0.000484730	0	0.003082583	0.015743008
$\mathbf{X}_C$	0.005712382	0.003082583	0	0.005572539
$\mathbf{X}_D$	0.020888985	0.015743008	0.005572539	0