Proceedings of the 5th WSEAS International Conference on Signal Processing, Istanbul, Turkey, May 27-29, 2006 (pp191-194)

# Discriminant Functions and Multi-Resolution Analysis (MRA) for Disease Detection

**Omar Mohd. Rijal*, Norliza Mohd. Noor†, Amran Hussin*, Ong Ee Ling***
* Institute of Mathematical Science, University of Malaya
†Diploma Program Studies, Universiti Teknologi Malaysia

*Abstract:*- Problems associated with the detection of diseases in their early stage are well known when using chest radiograph images. A graphical method involving wavelet coefficients as the feature vector (WFV) has been proposed for the detection and discrimination of Mycobacterium Tuberculosis (MTB) and lung cancer (LC). In a pilot study confirmed cases showing no complications (for example a section of the lung filled with water) were studied. Further, in the pilot study the feature vector were compressed for simpler data management. However discrimination using the compressed WFV was developed to handle small sample situations. In this paper, an alternative method for larger samples sizes was employed. A control group was used to calculate the parameters of the Linear Discriminant Function (LDF( $\underline{x}$ )) and the Quadratic Discriminant Function, (QDF( $\underline{x}$ )). A separate test group was then used to calculate misclassification probabilities. The feature vector selected, that is vector $\underline{x}$ , were the average and detail vectors from the MRA of the WFV. The technique developed here allows misclassified cases to be reclassified correctly, a facility not provided for in earlier techniques.

*Key-Words*:- Chest Radiograph, Mycobacterium Tuberculosis, Lung Cancer, Wavelet, Multi-Resolution Analysis, Discriminant Functions, Misclassification Probabilities.

## 1 Introduction

Despite rapid advances in medical imaging technology [1], [2], the conventional chest radiograph film is still an important ingredient in the diagnostic process. It is well-known [3] that mainly experienced medical officers are capable of accurately detecting MTB from chest radiographs whilst fewer doctors are capable of detecting LC in their early stage.

This study continues the work of discriminating MTB and LC given in [4]. The infected areas on digitized chest radiograph images of a confirmed MTB patient (or LC patient) was manually selected under the advice of a medical expert [3]. In a chosen infected area, vertical lines of pixels (called line profiles) were in turn selected. Thirty line profiles were obtained; see Figure 1 and Figure 2.

## 2 Multi-Resolution Analysis

Each line profile, say $\underline{x}_j$, is treated as a signal, and an MRA is performed on $\underline{x}_j$ using Daubechies 4 wavelet [5] and Daubechies 4 scaling signals. In other words

$$\underline{x}_j = A^1 + D^1 \quad (1^{\text{st}} \text{ level MRA})$$

where $A^1$ is called the first average signal and $D^1$ is called the first detail signal (following the terminology and notation of Walker [6]). For a $h^{\text{th}}$ level MRA we have;

$$\underline{x}_j = \underline{A}_j{}^h + \underline{D}_j{}^h + \cdots + \underline{D}_j{}^2 + \underline{D}_j{}^1$$

In this study, each signal $\underline{x}_j$ is regarded as being made up of 4 components; viz $\{\underline{A}^1, \underline{D}^1, \underline{D}^2, \underline{D}^3\}$. Since we have 30 line profiles (per patient) we have 30 sets of $\{\underline{A}^1, \underline{D}^1, \underline{D}^2, \underline{D}^3\}$. For patient –*k*, the average of each component is studied, namely $\{\overline{\underline{A}}_k{}^1, \overline{\underline{D}}_k{}^1, \overline{\underline{D}}_k{}^2, \overline{\underline{D}}_k{}^3\}$ , see [7]. Comparison between patient-*k* and patient –*m* may be done by comparing;

$\overline{\underline{A}}_k{}^1$ and $\overline{\underline{A}}_m{}^1$
or $\overline{\underline{D}}_k{}^1$ and $\overline{\underline{D}}_m{}^1$
or $\overline{\underline{D}}_k{}^2$ and $\overline{\underline{D}}_m{}^2$
or $\overline{\underline{D}}_k{}^3$ and $\overline{\underline{D}}_m{}^3$

## 3 A Graphical Method

It is difficult to compare large numbers of 26-dimensional vectors. Instead, each signal is reduced to a 2 dimensional curve called the Andrew's Curve [8]. Andrew [8] 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

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

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 average or detail signal) will appear as a harmonic curve drawn in two dimensions. For a given *t*-values, the distance between two curves $\{t, f_{\underline{x}_r}(t)\}$ and $\{t, f_{\underline{y}_r}(t)\}$ is proportional to the square of the Euclidean distance between $\underline{x}$ and $\underline{y}$. This property allows clustering [9] of a given set of vectors.

## 4 Data Sets

Following the notation of section 2, each patient may be represented by four signals, for example $\{\overline{\underline{A}}_j^{\,1}, \overline{\underline{D}}_j^{\,1}, \overline{\underline{D}}_j^{\,2}, \overline{\underline{D}}_j^{\,3}\}$ for patient –*j*. Given *n* patients, the subscript *j* take the values of $j = 1,2,\cdots,n$. The *n* patients will be represented by the following data sets;

Data Set (1) = $\{\overline{\underline{A}}_1^1, \overline{\underline{A}}_2^1, \cdots, \overline{\underline{A}}_n^1\}$

Data Set (2) = $\{\overline{\underline{D}}_1^1, \overline{\underline{D}}_2^1, \cdots, \overline{\underline{D}}_n^1\}$

Data Set (3) = $\{\overline{\underline{D}}_1^2, \overline{\underline{D}}_2^2, \cdots, \overline{\underline{D}}_n^2\}$

Data Set (4) = $\{\overline{\underline{D}}_1^3, \overline{\underline{D}}_2^3, \cdots, \overline{\underline{D}}_n^3\}$

## 5 Discriminant Using Data Set (*j*); *j* = 1, 2, 3, 4

Without loss of generality, consider the discrimination procedure using Data Set (1). Suppose we have *n*-confirmed MTB patients, and *w*-confirmed LC patient. We then consider the following two sets of vector $\{\overline{\underline{A}}_1^1, \overline{\underline{A}}_2^1, \cdots, \overline{\underline{A}}_n^1\}$ and $\{\overline{\underline{A}}_{n+1}^1, \overline{\underline{A}}_{n+2}^1, \cdots, \overline{\underline{A}}_{n+w}^1\}$ where the first set is derived from the MTB patients and the second from the cancer patients. Henceforth n-Andrew's Curve of $\{\overline{\underline{A}}_1^1, \overline{\underline{A}}_2^1, \cdots, \overline{\underline{A}}_n^1\}$ will represent MTB patients, and *w*-Andrew's Curve of $\{\overline{\underline{A}}_{n+1}^1, \overline{\underline{A}}_{n+2}^1, \cdots, \overline{\underline{A}}_{n+w}^1\}$ represent lung cancer patients.

Figure 3 shows the Andrew's plots for the control group of 40 MTB patients and 30 LC patients. At a given t-value ($t_0$) where graphs shows high separation we have

$\{f_{\overline{\underline{A}}_1^1}(t_0), f_{\overline{\underline{A}}_2^1}(t_0), \cdots, f_{\overline{\underline{A}}_n^1}(t_0)\}$ (MTB sample)

and $\{f_{\overline{\underline{A}}_{n+1}^1}(t_0), f_{\overline{\underline{A}}_{n+2}^1}(t_0), \cdots, f_{\overline{\underline{A}}_{n+w}^1}(t_0)\}$ (LC sample). Both of these sets were shown to be univariate normal (using the Kolmogorov-Smirnov test, [10]). As such we define $g_1(\overline{\underline{A}}^1)$ as the normal probability distribution for MTB sample and $g_2(\overline{\underline{A}}^1)$ for LC sample. Let

$$G(\overline{\underline{A}}^1) = \ln \frac{g_1(\overline{\underline{A}}^1)}{g_2(\overline{\underline{A}}^1)}$$

When the variances for $g_1(\overline{\underline{A}}^1)$ and $g_2(\overline{\underline{A}}^1)$ are equal, $G(\overline{\underline{A}}^1)$ becomes the LDF($\overline{\underline{A}}^{\,1}$), otherwise it is called the QDF($\overline{\underline{A}}^{\,1}$) [11].

If $G(\overline{\underline{A}}^1) > 0$ the vector corresponding to $\overline{\underline{A}}^{\,1}$ (from test set) belong to group 1 (MTB), otherwise it belongs to group 2 (LC).

The above steps performed the discrimination procedure for MTB and LC using the first average signal from where the classification probability may be calculated.

If all the above is repeated for Data Set (2), we will then perform the discrimination procedure for the first detail signal.

In total we performed the discrimination 4 times. Suppose the test set has *q* patients such that the first discrimination procedure correctly classifies $q_1$ patients ($(q_1 < q)$. Further suppose $q_j$ patients are correctly classified using the $j^{\text{th}}$ classification procedure therefore the probability of correct classification is $\dfrac{q_1 + q_2 + q_3 + q_4}{q}$.

## 6 Choice of *t*-values for Discrimination

Starting at $t$ = 0, with increment of 0.01 we calculate (for each *t*-value)

$$D(t) = \max_{\underline{x}} f_{\underline{x}}(t) - \min_{\underline{y}} f_{\underline{y}}(t)$$

assuming $f_{\underline{x}}(t)$ are greater than $f_{\underline{y}}(t)$ for all $\underline{x}$ and $\underline{y}$ . D(t) is calculated in an interval of length 0.5 for average signals and a length of 0.3 for detail signals, see Figure 3 and Figure 4. Finally in each interval the value t =to is chosen such that $D(t_0) = \max_t D(t)$ .

As a rule when $G(\underline{x}) > 0$ where $\underline{x}$ is either the average or detail signals for **all selected t-values** then the patient is said to be an MTB patient (otherwise a LC patient).

## 7 Result and Discussion

The average line profiles was treated as a signal (WFV) which was in turn transform to the Andrew's Curve. A comparison of averaged signals is equivalent to a comparison of their corresponding Andrew's Curves.

At a given *t*-value, two samples of observation were obtained from the Andrew's Curve. Each sample from either MTB patients or LC patients were tested and found to be normally distributed. This allowed the use of the LDF(.) for discriminating the average signals and the QDF(.) for discriminating the detail signals. The results of correct classification are given in Table 1.

In summary the LC patients showed better classification rate. This is most likely due to the more complicated configuration (for example shape, size and area) of the MTB image.

Further using the signals A1, D1, D2 and D3 consecutively, we have a procedure that can reclassify rejected observations.

Finally, the LDF(.) and QDF(.) were the best Discriminant Functions to be used because the data were normally distributed. Even so there were still misclassified cases which may be due to the image with medical conditions such as "cavitating opacity", or "cavitating lesion", [3].

## 8 Acknowledgement

We would like to acknowledge the contribution from:

(i) Datin Dr. Hjh Aziah Mahyuddin and Dr. Azwayati Abas, The Respiratory Unit, Kuala Lumpur General Hospital,and

(ii) Dr Hamidah Shaban, Selangor Medical Centre.

This research was funded under an IRPA grant from the Ministry of Science, Technology and the Environment and Universiti Teknologi Malaysia.

***References***

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

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

[3] Shaban, H, Medical Consultant for Respiratory Disease, Private Conversation, Institute of Respiratory Disease, Kuala Lumpur Hospital.

[4] Noor, N. M., Rijal, O. M., Shaban, H. and E. L. Ong, "Discrimination Between Two lung diseases Using Chest Radiographs" Proceedings of the 27th Annual international Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 05), Shanghai, 2005.

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

[6] Walker, J.S., A Primer on Wavelets and their Scientific Applications, Chapman and Hall, CRC Press, 1999.

[7] 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.

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

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

[10]Conover W. J., Practical Nonparametric Statistics, John Wiley, 1971.

[11]Lachenbruch, P. A., Discriminant Analysis, Hafner Press, 1975.

Table 1 Discrimination Result from using Data Set(*j*).

| Group | | Correct Classification |
|---|---|---|
| **MTB** | A1 | 22 |
| | D1 | 11 |
| | D2 | 6 |
| | D3 | 0 |
| | TOTAL | 39/50 |
| | PERCENTAGE | 78% |
| **LC** | A1 | 19 |
| | D1 | 3 |
| | D2 | 1 |
| | D3 | 0 |
| | TOTAL | 23/26 |
| | PERCENTAGE | 88% |

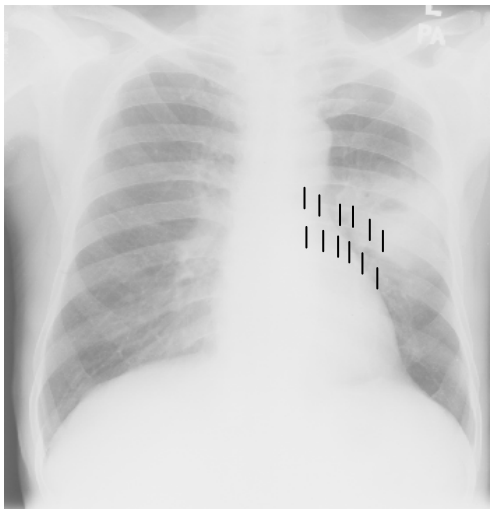

Figure 1 Confirmed lung cancer (left lung) chest X-ray.

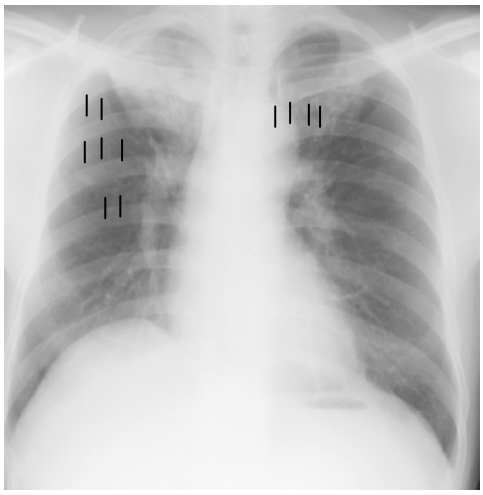

Figure 2 Confirmed MTB chest X-ray.

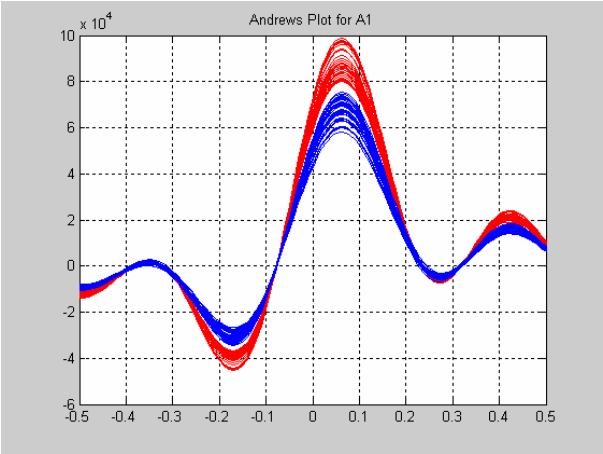


Figure 3 Andrew's Curve for first average signal of 40 confirmed MTB cases (blue line) and 30 confirmed LC cases (red line).

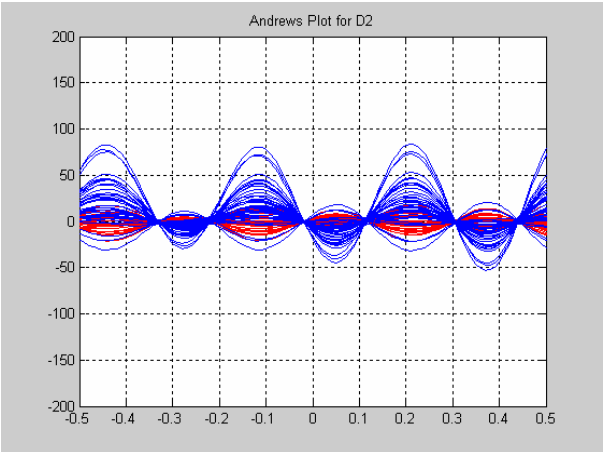


Figure 4 Andrew's Curve for second detail signal of 40 confirmed MTB cases (blue line) and 30 confirmed LC cases (red line).