Investigation of biomaterial classification using T-rays

Chin Cheng Te, Bradley Ferguson and Derek Abbott

Centre for Biomedical Engineering and Department of Electrical and Electronic Engineering,
The University of Adelaide, SA 5005, Australia

ABSTRACT

The use of terahertz pulses (T-rays) for imaging has created a wide range of new applications. This paper investigates a number of techniques for optimally classifying terahertz data. Specifically we consider statistical pattern classification methods.

A goal of this research is to implement a classifier as for classifying biomaterials. The objective is to train a classifier using THz images of known materials and then to use the classifier to identify the materials present in unknown images. Potential applications in security systems for airports, customs, and post offices are significant, because we can actually identify the material inside the package without opening it, based on a material's broadband frequency signature.

Keywords: Terahertz detection, T-rays, Statistical pattern classification

1. INTRODUCTION

"T-rays" – that is, terahertz (10^{12} Hz = 1 THz) electromagnetic pulses have a frequency range between microwaves and infrared radiation, and they offer scientists a new way to inspect, measure and analyze a diverse range of substances and materials. Due to the specialized femtosecond laser equipment needed, commercial applications were difficult to foresee before 1994.[1] significant recent progress has been made due to the present availability of semiconductor-diode-pumped solid-state lasers.[1] Moreover, the advances in terahertz transmitter, receiver, signal processing techniques and improved optical designs have also boosted up the commercial prospects of T-ray imaging.

One of the most promising applications of T-ray technology is package inspection. Laser pulses each lasting only 100 femtoseconds are used to generate electromagnetic T-ray pulses and then transmitted through various objects. By studying and analyzing the transmitted T-ray radiation through an object, we can actually distinguish materials by measuring the amounts of optical delay and absorption, as a function of frequency, from a T-ray receiver or detector. The advantage of T-rays, over say IR spectroscopy, is that THz frequencies correspond to large-scale molecular motions (eg. conformal states and 'breathing modes' for example) as opposed to the stretching resonances of, say, carbon-carbon bonds in the case of infrared (IR). The bottom line is that THz frequencies give information about the *whole* molecule, whereas higher frequencies pick up resonances in smaller atomic groups and bonds. This gives rise to a number of exciting T-ray applications from tagless detection of thin films in proteomic/genomic biochips through to, say, anthrax detection for security against bioterrorism. Another feature in support of biomaterial imaging applications of T-rays, is that microwave techniques suffer from poorer spatial resolution, whereas higher frequencies (eg. optical) suffer from Rayleigh scattering. T-rays represent the idea trade-off between these two extrema.

Digital signal-processing units can processes the T-ray data and then display a particular image on the computer screen. Example images obtained by processing T-ray data are shown in Figure 1. Non-polar, dry substances, such as cardboard and plastics, are transparent to terahertz radiation.[2] Molecular resonances occur in the terahertz band, so we can obtain richer information than by using other techniques. As T-rays are a low-energy non-ionizing radiation, for some applications, they might some day replace X-rays due to greater safety.

Biomedical Applications of Micro- and Nanoengineering, Dan V. Nicolau, Abraham P. Lee, Editors,
Proceedings of SPIE Vol. 4937 (2002) © 2002 SPIE · 0277-786X/02/$15.00

By using T-rays for package inspection, we can identify the material inside the package without opening it. This could compliment the existing security systems for airports, customs control, and post offices. Various types of methods can be used to analyze and classify the data obtained from a T-ray detector. An important problem is that of extracting features from the high dimensional THz data. An X-Y scatter plot provides an easy method of visualizing the success of different features in discriminating between different materials. In this paper we have investigated a number of different feature extraction methods and trained classifiers using the chosen features. THz images several materials including bacterial spores and common powders (see Figure 1) were used to train and test the classifier and the accuracy of the different feature extraction methods were compared.

This paper describes a method for biomaterials classification using T-rays. Section 2 provides an overview of pattern classification and model of classification that is used here. Section 3 discusses the methodology for implementing a classifier. Section 4 discusses the classification accuracy and compares this with a method based on feature selection. Section 5 summarizes this work and presents an overview of the future work planned in this developing field.

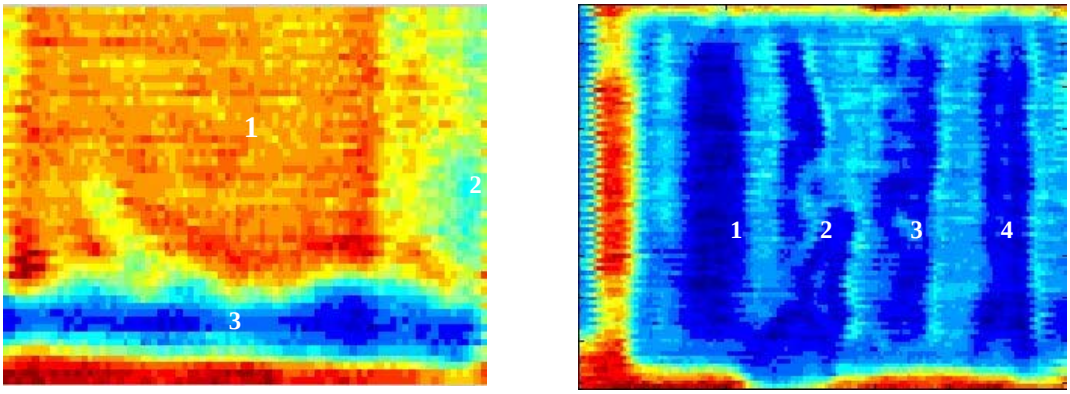


Figure 1. On the left-hand side, the image shows anthrax inside an envelope with a tape seal on it (1. Paper envelope; 2. Adhesive tape; 3. Bacillus spores). On the right-hand side, the image shows five different materials inside an envelope (1. Wheat flour; 2. Table salt; 3. Baking soda; 4. Chinese five-spice).

2. MODEL FOR STATISTICAL PATTERN CLASSIFICATION

Pattern classification or pattern recognition is the act of taking in raw data and taking an action based on the "category" of the pattern.[3] The demand for pattern classification systems has sharply increased due to the newly emerging applications in the areas of data mining, web searching, and chemometrics. Table 1 shows some examples of pattern classification applications.

The purpose of pattern recognition is to build a classifier that can take in some data and automatically produce an answer as to which class the data belongs to. Before implementing a classifier, basically the design of a classifier essentially involves the following three aspects: 1. Data gathering and pre-processing, 2. Feature selection or extraction, and 3. Decision making (method of classification). [3,4]

Suitable features are selected from the raw data. The choice of features is strongly based on the type of data or pattern. When the types of the data are speech, audio or similar waveforms, it is not possible to pick out some suitable features by just observing their original pattern. Therefore, a mathematical transformation on those data sets is needed. By using the Fourier or wavelet transform methods we can exploit some of their mathematical features.

Feature selection is normally based on the application or the input data type and generally falls in one of the following three classes[5]: 1. Mathematical features – Transform based features. Examples: Fourier Transform

coefficient, Wavelet Transform coefficient, Discrete Cosine Transform 2. Physical features – Intuitive Example: Peak value, rise time and 3. Model based – The data are modeled as solutions to some equations based on the model and from that we are able to choose the best features (parameters). Examples: FIR filter, IIR filter etc.

After the feature selection, one type of classifier is needed for analyzing the data. There are various types or models of classifier that are available. In this section, some appropriate methods will be introduced.

There are four well-known approaches for implement a classifier (See Table 2). These four models are not necessarily independent, because sometimes the same classifier method exists among them[4, 5]. In this paper, focus on the statistical approach. Many strategies are used to design a classifier in statistical pattern classification and basically it can be divided into two groups: parametric and non-parametric methods.

In the parametric method, one specifies a general formula for the probability distribution of the observation vectors for each class.[3] This is a classical pattern recognition technique based on linear discriminant functions; it uses the samples to estimate the values of parameters of the classifier and assumes that the forms of the discriminant functions are known.

For non-parametric methods, the user is not required to specify a probability distribution in advance. An example of a non-parametric method is nearest neighbor classification.[3] In this method, the entire training set is retained in memory. When a new case must be classified, the training set is searched for the example that is closest, according to some predefined metric. The new case is classified according to the distance to the nearest training example. [3]

Many approaches exist, this paper is not exhaustive and is only intended to give initial results. In this paper, we try to determine the most suitable classifier for our case. Figure 2 shows a simple block diagram to explain the approached to implement a classifier with a model of statistical classification.

TABLE 1: Examples of Pattern Classification Applications[5]

Problem Domain	Application	Input Pattern	Pattern Classes
Bioinformatics	Sequence analysis	DNA/Protein sequence	Known types of genes/patterns
Data mining	Searching for meaningful patterns	Points in multi-dimensional space	Compact and well-separated clusters
Document image analysis	Reading machine for the blind	Document image	Alphanumeric characters, words
Industrial automation	Printed circuit board inspection	Intensity or range image	Defective/ non-defective nature of product
Multimedia database retrieval	Internet search	Video clip	Video genres (e.g., action, dialogue, etc.)
Biometric recognition	Personal identification	Face, iris, fingerprint	Authorized users for access control
Speech recognition	Telephone directory enquiry without operator assistance	Speech waveform	Spoken words

Table 2: Pattern Classification Models[5]

Approach	Representation	Recognition Function	Typical Criterion
Template matching	Samples, pixels, curves	Correlation, distance measure	Classification error
Statistical	Features	Discriminant function	Classification error
Syntactic or structural	Primitives	Rules, grammar	Acceptance error
Neural networks	Samples, pixels, features	Network function	Mean square error

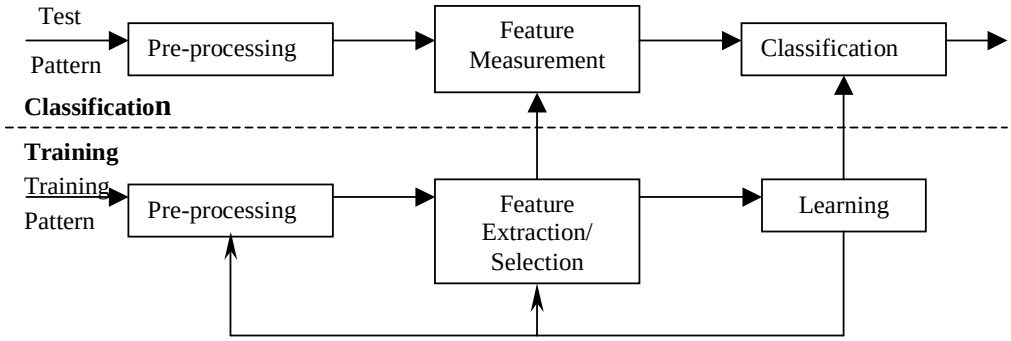


Figure 2 Model for statistical pattern classification. The system consisted of two parts that are training (learning) and classification (testing). At first the training data are input into pre-processing for filtering out some noise or redundant information, after that some suitable features are extracted from the data by applying some appropriate methods and then the classifier is trained to classify the feature space. The feedback path allows the user to optimise the process of pre-processing and feature extraction. In the classification mode, the unknown material or test sample is inputted into the classifier for classification, which is based on the previous consideration or method in the training mode. [3,5]

3.METHODOLOGY

In this paper, T-ray data from different materials (paper, bacterial spores, soda, salt, flour and Chinese seasoning) were used to train a classifier. The T-ray raw data is shown in Figure 3. The data are in the form of signal waveforms in the time domain, which were acquired using a T-ray imaging system.[6]

Before processing the data, we can pre-process it by filtering out some of its noise. Also further preprocessing may be used to either cleanup the image or to generate a compact representation of the pattern. Segmentation is also one of the techniques to pre-process the data, so clearer features of that data would be observed. Pre-processing also helps to reduce some redundant features that would make the task more complicated.

There is an important step in the selecting of distinguishing features, in order to minimize the error rate of classification. A limited, but important features set, simplifies both the pattern representation and the classification processes that are built on the selected representation. As a result, the classifier will be faster and use less memory. The decisions for selecting features are based on: 1. Finding features that are simple to

extract; 2. Invariant to irrelevant transformations; 3. Insensitive to noise and 4. Useful for discriminating patterns in different categories.[3]

3.1 PHYSICAL FEATURES (INTUITIVE)

By simply looking at the raw data displayed in Figure 3, we can see seven groups of pulses overlapping with one another. The straightforward way to select features, in order to classify these seven groups of data, is by choosing the pulse peak value and its corresponding time value. After choosing the peak value and its corresponding timing as the data to train the classifier, an XY scatter plot can be generated to see the performance of the classifier before we actually using it as the training samples. The result is shown in Figure 3.

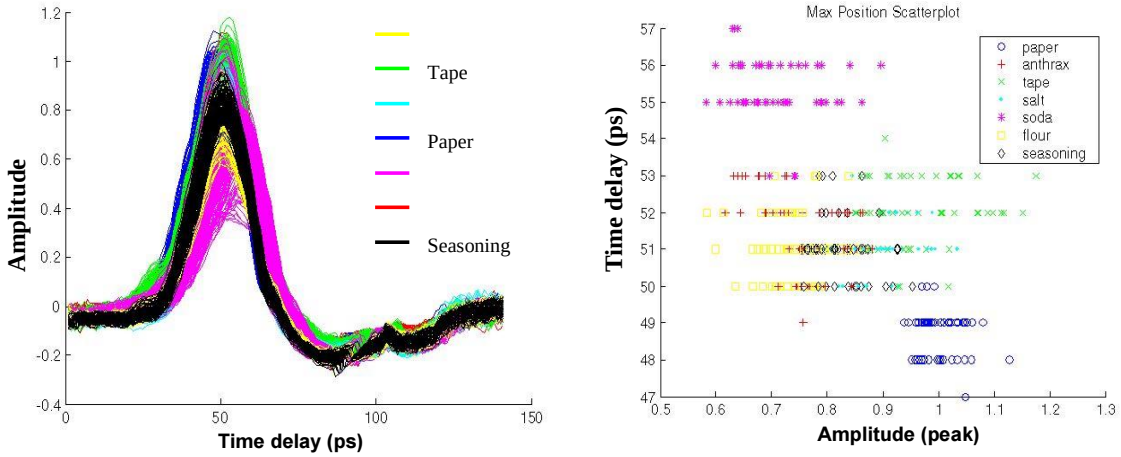


Figure 3. The graph on left-hand side shows the seven groups of raw data, which are flour, tape, salt, paper, soda, bacillus spores and seasoning. On the right-hand side we observe that there are seven groups of data, which cluster according to the selected features. Many points are overlapping and only paper and soda are well separated, so we can predict that the result of this classification is not good if we use these features to train the classifier.

3.2 MODEL BASED

We used parametric identification methods to create a mathematical model for the materials' THz responses based on the training data. Basically it is a matter of finding (by numerical search) those numerical values of the parameters from some given models that give the best agreement between the model's (simulated or predicted) output and the measured one.[7] Most common given models are difference equations descriptions, such as AR and ARX models, as well as all types of linear state-space models. Here, we only concern ourselves with the ARX model.

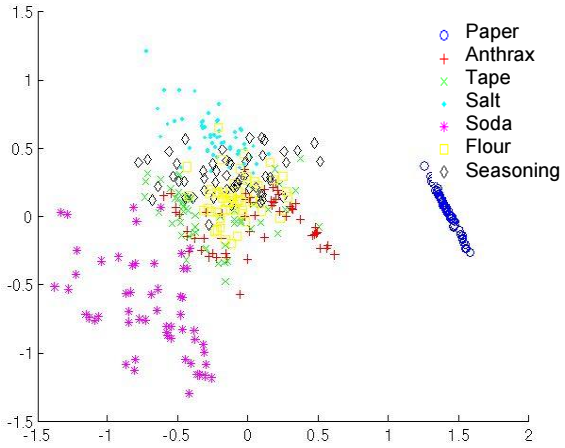
ARX is a given model containing the structure of simple linear difference equation that relates the input $u(t)$ to the output $y(t)$ as follows[7]:

$$y(t)+a_1y(t-1)+\ldots+a_{na}y(t-na) = b_1u(t-nk)+\ldots+b_{nb}u(t-nk-nb+1) \quad (1)$$

The structure is defined by the three integers *na*, *nb* and *nk*. *na* is the number of poles, *nb+1* is the number of zeros, and *nk* is the pure time delay (the dead time) in the system. Those parameters are estimated based on the

least squares method, which minimizes the sum of squares of the right-hand side minus the left-hand side of the expression above, with respect to *a* and *b*.[7]

After the order of ARX is determined, some parameters could be extracted by applying some simple Matlab code. Then a XY scatter graph can be plotted to predict which order of ARX would result in a better classification. Some XY scatter graphs with different order of ARX are shown in Figure 4.



ARX order [0 3 1]

Figure 4. The overlapping circumstance is decreased by using the model based method when compare to figure 3. Data points are much more scatter, this make the job of classifier much more easier.

3.3 MATHEMATICAL FEATURES

3.3.1 WAVELET TRANSFORM

A wavelet is a waveform of limited duration that has an average value of zero.[8,9] It is useful in the analysis of non-stationary signals and in image processing, because wavelets tend to be irregular, asymmetric and localized in time and frequency.[8, 10] Besides that, signals with sharp changes or a pulsed shape might be better analysed with an irregular wavelet than with a smooth sinusoid, by intuitively looking at the wave shape of sine wave and wavelet.[8, 11]

The purpose of the wavelet transform in our context is to extract the wavelet coefficients out from the raw data as the input features for a classifier. All we are concerned with is the accuracy of our classifier. This can be done by using the multi-level one dimensional wavelet decomposition in Matlab. We only consider on the Symlet family with an order 2, because the other families with the same order would result a similar accuracy on the classifier (see Section 4). Figure 5 shows some scatter plots resulting from the extracted wavelet coefficients by using different wavelet families.

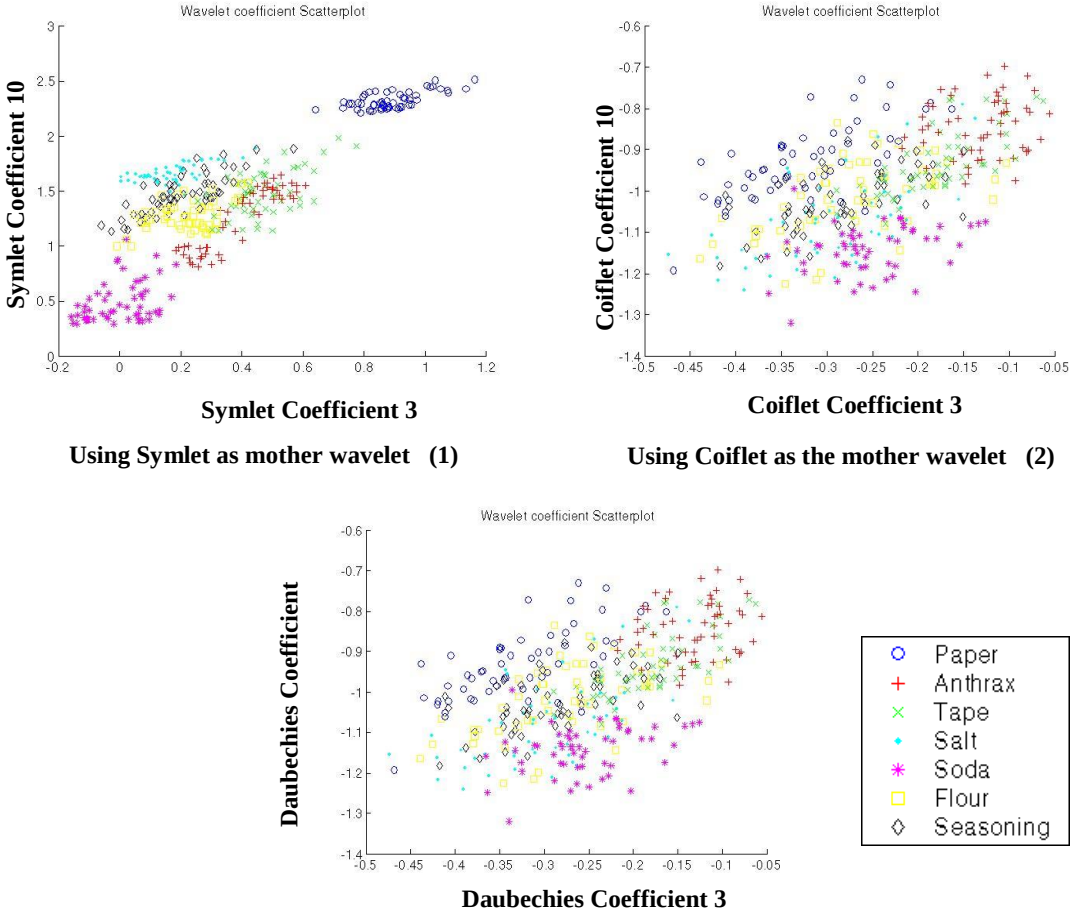


Figure 5. (1) Shows the better classification result. (2) and (3) show very similar performance because they have the same wavelet coefficients. From the result above, we can see that, the Symlet with the order 2 will result a more accurate classifier. We using coefficient 3 and 10 are because it illustrates a much better-looking scatter plot.

3.3.2 FAST FOURIER TRANSFORM (FFT)

Applications of Fourier analysis for non-periodic discrete signals results in the discrete Fourier transform representation. The discrete-time Fourier transform pair is defined by[9, 12]

$$x[n]=\frac{1}{2\pi}\int_{\Omega_0}^{\Omega_o+2\pi} X(\Omega)e^{i\Omega n}d\Omega \quad (2)$$

$$X(\Omega)=\sum_{n=-\infty}^{\infty} x[n]e^{-j\Omega n} \quad (3)$$

Where Ω_0 is an arbitrary real number. $X(\Omega)$ is called the discrete-time Fourier transform of $x[n]$ and is a continuous periodic function[7] with a period of 2π.

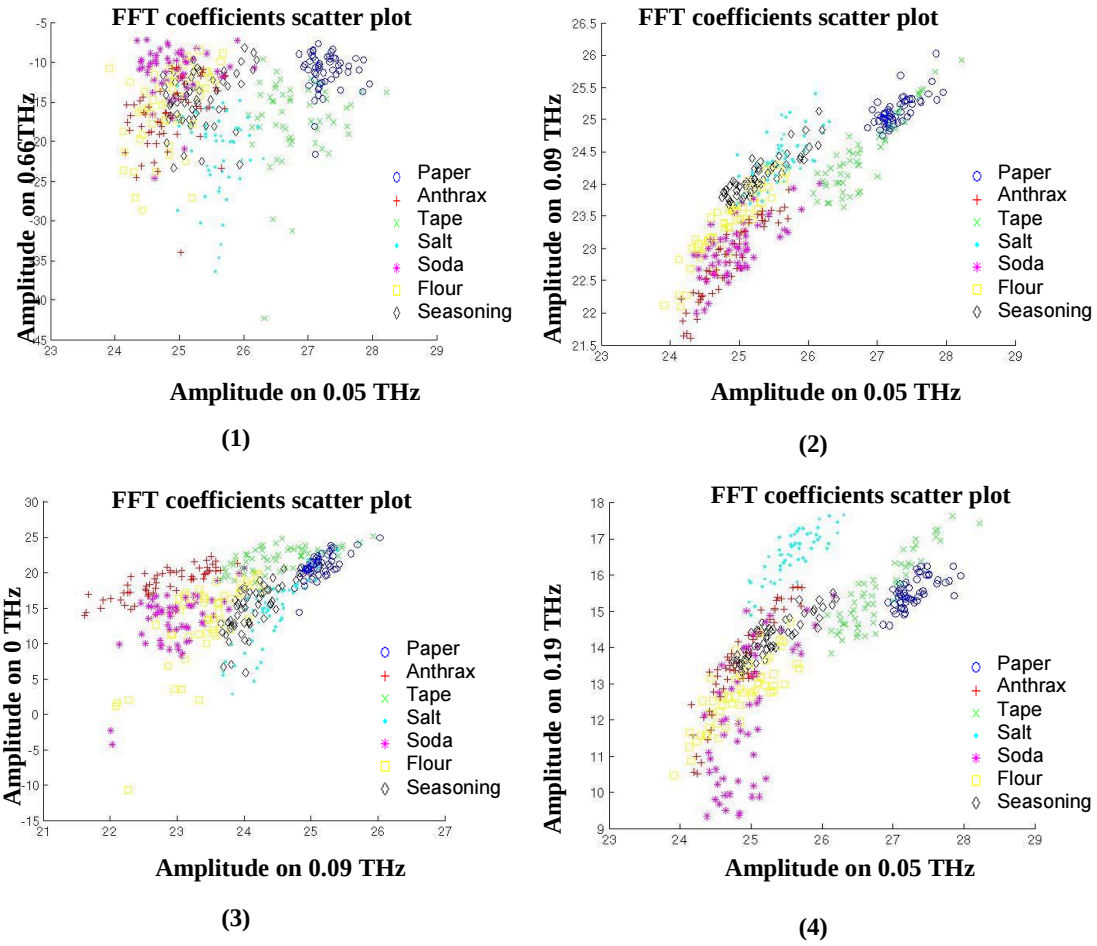


Figure 6. Plot (4) shows the best classification result among them, this is because the graph number (4) is more dispersed than the others and this will make the job of the classifier easier. The worst one is (1), due to the overlapping of the data points. The accuracy of classification (2) and (3) are quite similar, because the overlapping circumferences for them are equal likely. (Graph (1) is result from taking frequency components of 0.05,0.76,0.66, and 1.23(THz), for graph (2) is 0.05,1.51,0.09, and 0 (THz), for graph (3) is 0.09,0.14,0,0.47(THz) and the last one is graph (4) which is result from the frequency components of 1.84,0.05,0.19 and 0.14 (THz))

The FFT is the fast algorithm for computation of the DFT, although this algorithm is not a transform, it was

named the fast Fourier transform (FFT). [9,12] Figure 6 shows the scatter plot resulting from choosing particular frequency components out from the data after taking an FFT. The particular Fourier phases from the data after taking FFT are also considered as the input features for the classifier. The result of this is shown by using some scatter plots in Figure 7.

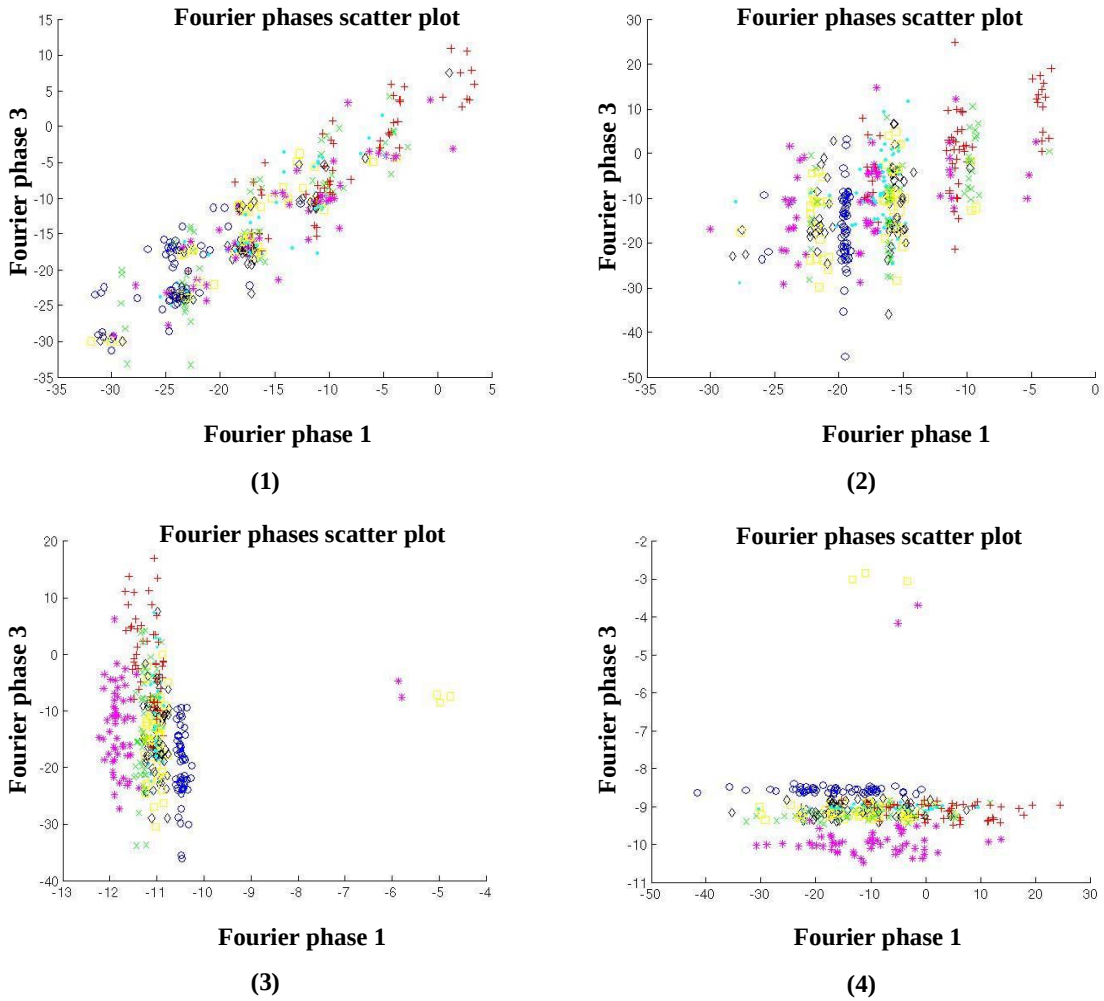


Figure 7. The worst result is (1) and the best one is on number (4). The results for (3) and (4) are similar. This result can also be compared with the result shown in Figure 6. From there, we can know that using the FFT coefficients as the input features of classifier is better than using the Fourier phase. (Graph (1) is result from taking the combination of Fourier phase that occurs on 1.04,1.18,1.28, and 0.19 (THz), for graph (2) it is 0.61,0.57,1.94, and 0.85 (THz), for graph (3) it is 0.24,1.65,1.42,and 0.05(THz), and for (4) it is 1.84,0.05,0.19 and 0.14 (THz)).

4. RESULTS FROM INVESTIGATION

Linear discriminant analysis is a very useful statistical tool. It takes into account the different variables of an object and works out which group the object most likely belongs to. Whether or not two or more groups are significantly different from each other with respect to the mean of particular variables.[13,14,15]

We use the Mahalanobis distance as a multivariate measure of the separation of a data set from a point in space. It is the criterion minimized in our linear discriminant analysis fro classification.[15, 16]

The Mahalonobis distance classifier was trained using 20 training pixels for each of the seven different powders considered. It was then tested using another 20 responses for each material. By substituting different

input features that are extracted based on the methods shown in Section 3, the feature extraction methods can be compared based on the accuracy of the resulting classifier. Different order ARX models were tested and the third order finite impulse response filter was found to be perform well, with only marginal improvement with increasing order. The classification accuracy of this filter was 78.75%. The ARX model was found to have difficulty differentiating between the adhesive tape and wheat flour materials. When these were omitted from the classification test the performance-improved markedly to 88% as shown in Table 3.

The frequency response of solids at THz frequencies are generally very complex and consist of a multitude of dense resonant frequencies resulting from intra and inter molecular modes. In general it is not possible to theoretically calculate frequencies of interest for identifying materials except in very simple cases such as gas spectroscopy. As a result we have adopted an iterative algorithm for identifying frequencies of interest. Our algorithm randomly selects a number of different frequencies. These frequency components are then used as the feature vector to train the classifier and the resulting classifier accuracy is measured. Over time this algorithm is capable of identifying suitable frequencies for material identification. This process can be used with either the Fourier amplitude or the phase of the Fourier coefficients as shown in Table 4 and Table 5 respectively.

The wavelet coefficients are extremely attractive tools for classifying THz responses for the reasons described in Section 3. A similar method to the Fourier analysis was used to identify the wavelet coefficients of interest. These were then used to train a classifier and the accuracy recorded as before. A number of different wavelet families were considered and the Symlet offered the highest accuracy of over 99%! These results are shown in Table 6. Table 7 then compares the classifier accuracy using each of the different feature extraction methods.

Table 3. Accuracy of classifier by using the coefficient from ARX model with order [0 3 1]

ARX order	Total number of input sample	Number of sample been classified correctly	Correct classified (%)
[0 3 1]	140	110	78.57
[0 3 1]	100	88	88.00
[0 3 1]	80	79	98.75

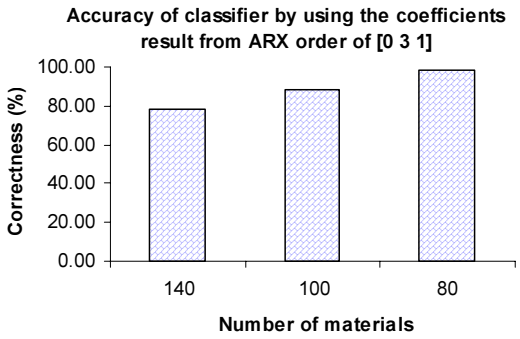


Table 4. The accuracy of a classifier that randomly selects the frequency components of T-ray data after FFT.

Different FFT combination (THz)	Total number of input sample	Number of sample been classified correctly	Correct Classified (%)
0.05,0.66,0.76,1.23	140	85	60.71
0,0.05,0.09,1.51	140	104	74.29
0,0.09,0.14,0.47	140	111	79.29
0.05,0.14,0.19,1.84	140	122	87.14

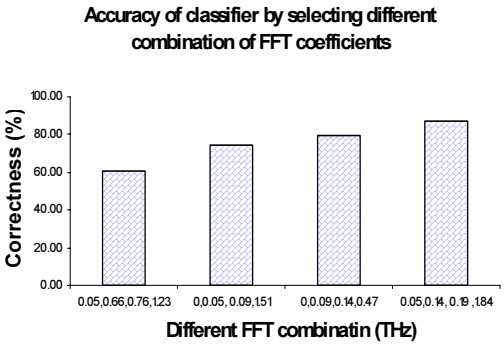


Table 5. Accuracy of a classifier that randomly selects the Fourier phase of T-ray data after FFT.

Different Fourier phase combination (THz)	Total number of input sample	Number of sample been classified correctly	Correct Classified (%)
0.19,1.04,1.18,1.28	140	59	42.14
0.57,0.85,0.61,1.94	140	74	52.86
0.05,0.24,1.42,1.65	140	88	62.86
0.05,0.14,0.19,1.84	140	108	77.14

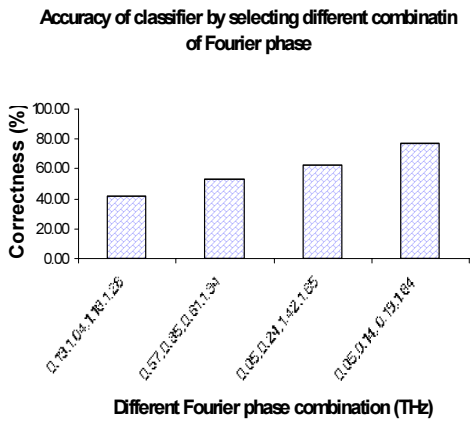


Table 6. Compares the accuracy of a classifier by using different wavelet families with order 2.

Wavelet family	Total number of input sample	Number of sample been classified correctly	Correct Classified (%)
Symlet	140	139	99.29
Coiflet	140	130	92.86
Daubechies	140	130	92.86

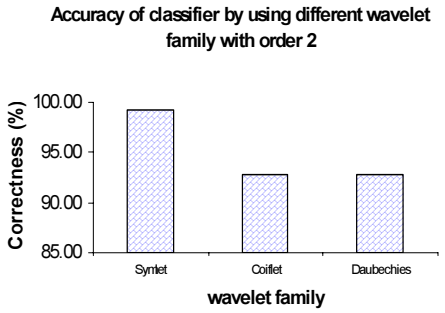
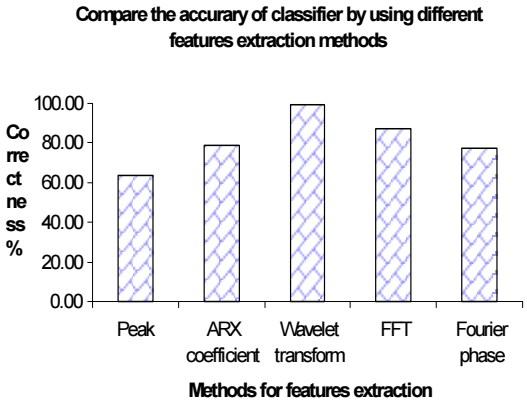


Table 7. The comparison of the accuracy of the classifier by using different feature extraction methods.

Features extraction Methods	Format	Total number of input sample	Number of sample been classified correctly	Correct Classified (%)
Physical features (Intuitive)	Peak	140	89	63.57
Model based	ARX coefficient	140	110	78.57
Mathematical features	Wavelet transform	140	139	99.29
	FFT	140	122	87.14
	Fourier phase	140	108	77.14



5. CONCLUSION

This article has investigated the performance of simple classifier by using different methods of feature extraction from T-ray data.

Feature extraction via the wavelet transform was found to be well suited for biomaterial classification using T-rays. This is primarily because signals with sharp changes or a pulsed shape tend to be more efficiently decomposed into wavelets. More signal power appears in the coefficients of interest.

The fast Fourier transform was also applied for feature extraction. The classifier accuracy that was achieved by this method is also acceptable (87.14%). This was achieved using the frequency components at 0.05, 0.14, 0.19 and 1.84 THz as features. This may indicate that these frequencies are important resonant frequencies of the materials considered. This is a dynamic field of research and much work remains. Besides using the linear discriminant analysis, other methods like partial least square algorithms (PLS), nearest mean classifier etc. will be investigated in the future.

ACKNOWLEDGMENTS

Funding from the ARC is gratefully acknowledged.

REFERENCES

1. D. M. Mittleman, R. H. Jacobsen, and M. C. Nuss, "T-ray Imaging," *IEEE J. Selected Topics in Quantum Electronics*, **Vol. 2,** No 3, pp. 679-692, 1996
2. B. Ferguson and D. Abbott, "Wavelet de-noising of optical terahertz pulse imaging data," *Fluctuation and Noise Letters*, **Vol. 1**, No. 2, pp. L65-L70, 2001.
3. R.O. Duda, P. E. Hart and D. G. Stork, *Pattern classification*, 2nd Edition, John Wiley & Sons, Inc., New York, 2001
4. R.O. Duda, and P. E. Hart, *Pattern classification and scene analysis*, John Wiley & Sons, Inc., New York, 1973.
5. A. K.Jain, R. P.W. Duin, and J. Mao, "Statistical Pattern Recognition: A Review," *IEEE Pattern Analysis and Machine Intelligence*, **Vol. 22**, No. 1, January 2000
6. B. Ferguson, S. Wang, D. Gray, D. Abbott and X. -C. Zhang, "Identification of biological tissue using chirped probe THz imaging," *Microelectronics Journal*, (To Appear).
7. L. Ljung, *System Identification Toolbox User's Guide version 5*, The MathWorks Inc., 2000
8. M. Misiti, Y. Misiti,G. Oppenheim and J-M. Poggi, "*Wavelet Toolbox User's Guide version 2.1*," The MathWorks Inc., 2000
9. C.H. Chen, *Signal Processing Handbook*, Dekker, New York, 1988
10. L. Prasad and S. S. Iyengar., *Wavelet analysis with applications to image processing*, CRC Press, 1997
11. B. W. Suter, *Multirate and wavelet signal processing*, Academic Press, San Diego, 1998
12. G. Bachman, E. Beckenstein, and L. Narici, "*Fourier and wavelet analysis*," Springer, New York, 1999
13. V. Cherkassky and F. Mulier, *Learning From Data*, John Wiley, 1998
14. S. Balakrishnama and A. Ganapathiraju, *Linear Discriminant Analysis- A Brief Tutorial*, Mississippi Sate University, 1998
15. "*Statistics Toolbox User's Guide version 3*," The MathWorks Inc., 2000
16. J. F. Hair, "*Multivariate data analysis with readings*,"4th Edition, Prentice-Hall, 1995.