Optical and Acoustical Methods in Science and Technology

Terahertz Frequency Domain Spectroscopy Identification System Based on Decision Trees

R. Ryniec*, P. Zagrajek and N. Pałka

Institute of Optoelectronics, Military University of Technology, S. Kaliskiego 2, 00-908 Warsaw, Poland

The application of pattern recognition methodology within chemistry, biology and other science domains, especially in security systems is becoming more and more important. Many classification algorithms are available in literature but decision trees are the most commonly exploited because of their ease of implementation and understanding in comparison to other classification algorithms. Decision trees are powerful and popular tools for classification and prediction. In contrast to neural networks, decision trees represent rules, which can readily be expressed so that humans can understand them or even directly use in a database. In this paper we present an algorithm of construction of decision trees and a classification rule extraction based on a logical relationship between attributes and a generalized decision function. Moreover, correctness and efficiency of the algorithm was experimentally validated in a terahertz system, where spectra of explosives were measured in reflection configuration.

PACS: 42.81.Bm, 42.81.Cn, 42.81.Dp

1. Introduction

The terrorist threats have been increasing worldwide during the last decades. Media reports about explosions have become common place, with the greatest threat emanating from suicide bombers in crowds and car bombs in traffic. The societal, economic, and political sphere are interested in having all technological options exhausted to prevent such attacks. So far, no stand-off detection devices are available that will detect potential assassins from a safe distance [1]. In the field, well-trained sniffer dogs are the best alternative for sensing explosives remotely, albeit at distances of no more than a few meters. Portal technologies and sampling detection systems are unsuited for stand-off detection. The terahertz (THz) region of electromagnetic spectrum offers an innovative sensing technique that provides information unavailable in other conventional methods. The use of T-rays, or terahertz radiation, to identify substances by their spectroscopic fingerprints is a rapidly moving field.

In the last few years, a number of researchers including our collaboration have assembled databases of terahertz (THz) time-domain spectroscopy (TDS) absorption and reflectance spectra from bulk explosives. While this was a necessary and important step in demonstrating the feasibility of THz TDS for explosives detection, the goal of our research is to develop system based on automatic recognition in real time at standoff distance. The dominant approach is presently terahertz time-domain spectroscopy. However, a key problem is that ambient water vapour is ubiquitous and the consequent water absorption distorts the T-ray pulses. Water molecules in the gas phase selectively absorb incident T-rays at discrete frequencies corresponding to their molecular rotational transitions. This situation, therefore, motivates the need for an optional alternative method for reducing these unwanted artefacts.

Ambient water vapour is commonly removed from the T-ray path by using a closed chamber during the measurement. Yet, in some applications, a closed chamber is not always feasible.

In the main body of this paper we described the methodology of computation of the populations variations of explosives and classification based on decision trees. The paper is organized as follows. Initially, the population of compounds is obtained by deformed "pure" spectra using the complex frequency response of water vapour modelled from a spectroscopic catalogue (software HITRAN). Then using decision tree for feature selection and classification was discussed.

2. Background

To identify potential suicide bombers effectively there should be met low false alarm ratio and detectability of wide range different explosive formulation requirements inter alia. Together, these performance requirements demonstrate clearly the technical challenge involved in developing suitable measuring systems for the stand-off detection of explosives. The methods used up to now still require approaching the suspicious object in order to perform the analysis with great risk for operator. A suitable

* corresponding author; e-mail: rryniec@wat.edu.pl

analytical instrument should be able to detect and identify at a stand-off distance the explosive materials with a reasonable level of confidence in order to offer real-time results maintaining a security distance for the operator. It has been recognized that laser-based spectroscopy is the only technique, which may be potentially capable to standoff detection of minimal amounts of explosive materials in real field scenarios [1]. After acquiring a measurement of terahertz pulse interacting with a sample, it is compared to a reference terahertz pulse (not interacting with the sample) and the material's optic properties are extracted. These properties (absorption, reflectance) manifest itself as absorption peaks (or change the reflectance) at specific frequencies.

However, conventional THz-TDS is dominated by manual analysis. The operator compares the peak positions to positions of known peaks to identify a sample. But, the reliance on human input is questionable because of few reasons [1]. The most important is, that for large data sets, the manual process can be slow. Performance automated identification systems of samples with terahertz spectrometry is a purpose of many research teams all over the world. In the literature [2] several attempts have been performed. These involved techniques such as linear correlation, the Mahalanobis distance, neural network, least-squares, and principal component analysis. Among these classification algorithms mentioned before decision tree is the most commonly used because of that it is easy to understand and implement [2–4]. Another problem is inflexibility to changes in environment. In practical detection system THz radiation is directed at a target, from which it is reflected back measured by a detector (Fig. 1).

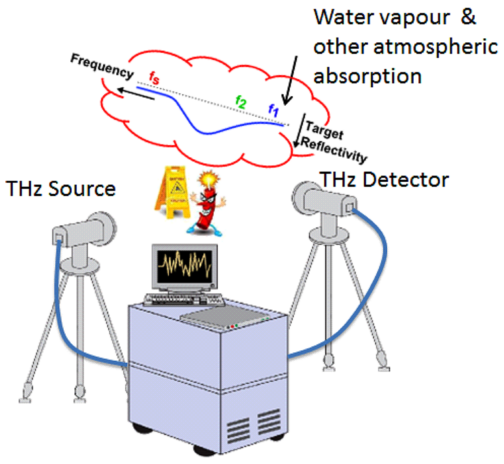

Fig. 1. Typical remote threat detection system geometry.

However, this radiation has to propagate through the atmosphere before reaching the target. It will then be reflected and scattered by possibly rough and irregularly shaped explosives before returning to the detector. When T-rays propagate through an atmosphere, fluctuations after the main pulse in the time domain are observed. So, T-ray spectroscopy of a sample, in open air, therefore often results in a combination of the sample's spectral features and water vapour resonances in the frequency domain. These effects are generally undesired, since they may mask critical spectroscopic data. Consequently, the target spectra must be sufficiently strong to be able to overcome the signal attenuation. Using only reduced information from spectra (the most informative set of frequencies and corresponding with them reflectance — called feature space), which not defeated by atmosphere attenuation is possible solution. Feature selection is significant step to improve existing algorithms for automated classification of explosives in stand-off THz systems. The decision tree both for obtaining a feature space and for classification examples is discussed.

3. Description of the approach

3.1. Acquiring data set

Time domain spectroscopy is commonly used technique in THz range. We have carried out measures of reflectance spectra $R(\omega)$ of five compounds in two reflection configurations — specular, where the sample is placed close to the detector with incidence and collection angle of the laser beam equal to 45°, and stand-off in the compartment purged with dry air at the distance 30 cm and incident angle equal to 5° [4]. The 20 ps wave form is converted to the frequency domain via the Fourier transform. The usable frequency band, used in our analysis, is $0.3 \div 2$ THz. For each material of interest, reflectance measurements are made in specific frequency range. Data $R(\omega)$ consist of pairs $\{(\omega_k, r_k)\}$, $k = 1, 2, \ldots, K$, where ω_k is a frequency and r_k is a measurement of reflectance at that frequency. Thus, we have $L = 45$ column vectors $(R(\omega)_1, R(\omega)_2, \ldots, R(\omega)_L)$, each has $N = 281$ frequency components and represents a THz reflectance spectrum (9 spectra per material — 3 spectra from pellet measured in specular geometry and 3 from a second, thicker pellet and 3 spectra from stand-off geometry [5, 6]). In order to get sample spectra the variation in measured reflectance due to propagation in free air, we used set of 11 numerical atmosphere transmission models $H(\omega) = \{(\omega_k, h_k)\}$, based on HITRAN software (different distance and humidity) to predict the shape of spectra $X(\omega) = \{(\omega_k, x_k)\}$ deformed by the atmosphere. A calculation of shape of spectra $X(\omega)$ of measured reflectance spectra $R(\omega)$ may be expressed as follows:

$$\boldsymbol{X}(\omega) = \boldsymbol{H}(\omega) \cdot \boldsymbol{R}(\omega). \tag{1}$$

Thus, we obtained 495 reflectance "measurement" vectors $X = \{X_1, X_2, \ldots, X_n\}$, $n \in \langle 1, 495 \rangle$, 99 spectra per materials, has $N = 281$ frequency components and represents THz reflectance spectra from RDX, sugar, salicylic acid, picric acid and para-aminobenzoic acid (PABA) (Fig. 2). Prior to applying the pattern recognition tool extensive pre-processing was performed. Samples vector X_n were normalized to remove sample to sample systematic variation usually associated with the total amount

of sample, by the formula

$$X_n = \sum_{i=1}^{\max} |X_p|. \tag{2}$$

Each element in the vector is divided by the constant obtained by the sum of the absolute value of all of the i entries in the vector X_p. These data sets were entries for algorithms for feature selection based on decision tree (Fig. 3) which is described in Sect. 3.3.

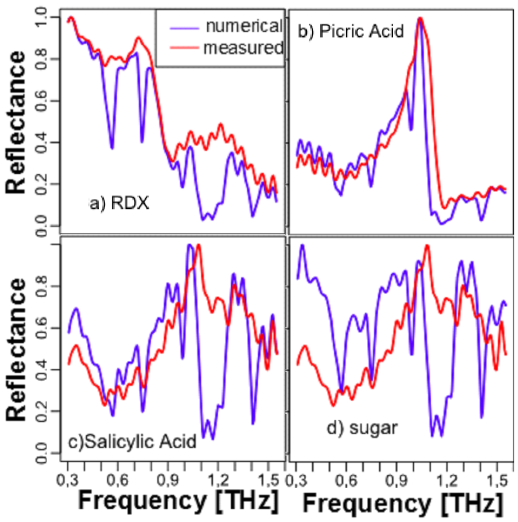

Fig. 2. Influence of water vapour on reflectance spectra: (a) RDX, (b) picric acid, (c) salicylic acid, (d) sugar.

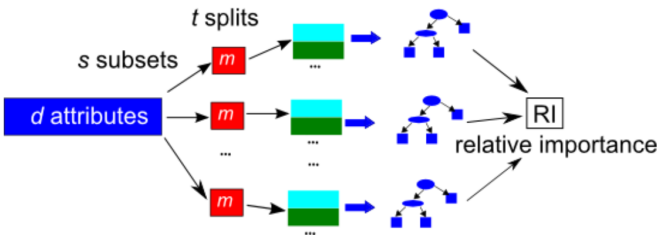

Fig. 3. Feature selection algorithm based on decision tree and random training data sets.

3.2. Properties of decision trees

Decision tree is characterized by few properties. The most interesting strengths of this method is that decision trees generate understandable rules (result depends on training set) without requiring much computation. Additionally, decision trees provide a clear indication of which fields are most important for prediction or classification but results depend on input data sets. Due to the unstable nature of decision trees, they are ideal for purpose of strengthening the classification algorithm. We can use it to combine outputs of many "weak" classifiers based on decision tree to produce a powerful "committee". This method called ensemble learning is one of the most powerful learning ideas [6].

3.3. Feature selection algorithm

The idea of "ensemble classifiers" was used to procedure both providing a better classification accuracy and reducing the cost of recognition by reducing the number of features that need to be collected. Random input selection was used in order to promote further diversity by selecting random s subsets of m-dimensional vectors to perform the learning task. The $m = \sqrt{d} = 281$ value was obtained in accordance with [6]. Random split into training and test subset was applied. The training data sets were used in building the classification mode, while test data records were used in validation of the model. A plot of exemplary result of classification model is found in Fig. 4. The $F2 < 0.985455$ represents classification rule which allowed to classify unknown reflectance spectra of compound as RDX and PABA or other compound of five class of materials. Attribute $F2$ correspond with second of 281-dimensional set of measure frequency.

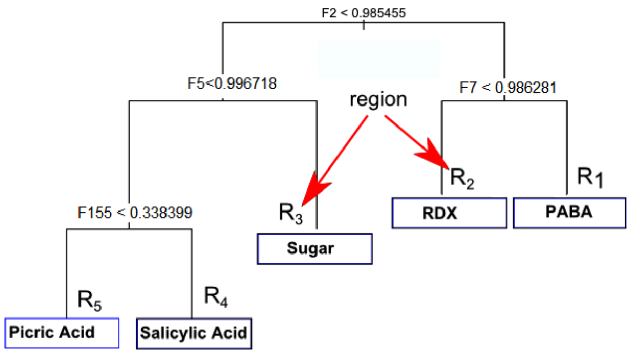

Fig. 4. Feature selection based on decision tree.

The classification rules $\{F2, F7, F5, F155\}$ shown above were obtained by the formula called deviance [7]:

$$Q(R_k) = -2 \sum_{s=1}^{u} N_s(k) \log_2 p(P_s|P_k), \tag{3}$$

where N_s — number of observation of class compounds $P_s, s \in \langle 1, u = 5\rangle$ in R_k classification region. The approach taken here allowed us to obtain the rank the most interesting, from the classification point of view, measure frequencies. We have chosen set of eighteen frequencies, which mostly attended in construction classification model with good prediction accuracy (Table I).

TABLE I

Set of frequencies for construction of feature space.

GHz	654.32	641.74	635.45	648.04	660.07	639.16
No.	1	2	3	4	5	6
GHz	905.99	893.34	666.91	899.7	880.82	887.12
No.	7	8	9	10	11	12
GHz	874.53	868.24	836.78	824.20	861.95	673.20
No.	13	14	15	16	17	18

3.4. Experimental setup

The classification of collected signals is often the most important step in detection and identification systems. It can be defined as assigned unknown class of compounds P of an observation X to data objects based on relationship between the data items with a pre-defined class label. To verify classification procedure the experimental setup was prepared. It allows measurements in real conditions with signal attenuated by water vapour. The laboratory setup based on subset of measure frequencies was built. The T-ray system was used with tuneable Virginia Diodes (VDI) source of radiation based on the Schottky diode achieving radiation power 1 mW–300 GHz to 1 nW –950 GHz, and pyroelectric radiometer cooperating with DSP Lock-In-software.

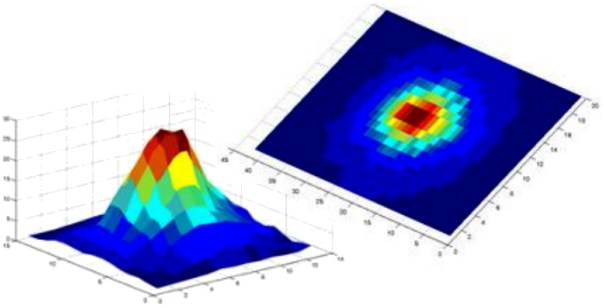

Fig. 5. Focused THz radiation beam.

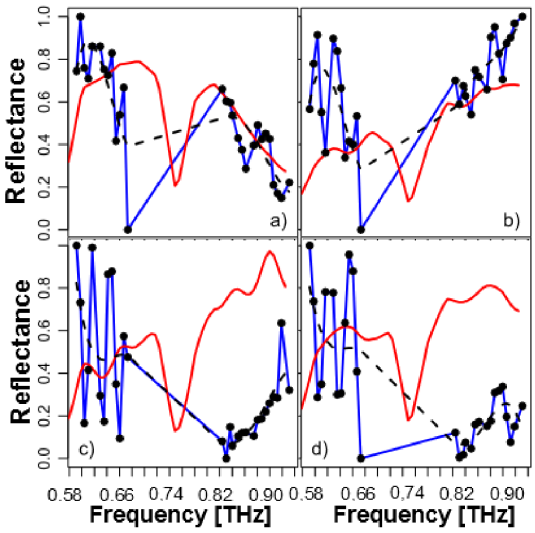

Fig. 6. Results of experimental setup: (a) RDX, (b) picric acid, (c) salicylic acid, (d) sugar (measure points — black, blue line — interpolated curve of spectra, red line — spectra obtained by TDS system).

The incident angle of THz beam onto the sample's surface is 45°. A polyethylene lens with 75 mm focal length was used to image the target onto a sensor (19.6 mm^2). To ensure the best acquisition quality system was calibrated before to get on sample and detector radiation beam focused into 19.6 mm^2 area. The power distribution of laser beam is shown in Fig. 5.

We have carried out measures for the samples as were used in TDS system. Then, reflectance was obtained. We used variation of measurements as a separate observation. Thus, we obtained data set containing 919 reflectance measurement vectors of 6 classes of compounds. A plot of results compared with TDS results is found in Fig. 6. As it is shown, we can find quite good agreement between measured data (blue) and predicted (red line).

4. Results and discussion

After preparation of "feature space" we have obtained collection of spectra based on experimental setup and real weather conditions. The number of observations of each class of compounds is shown in Table II. We additionally measured mixture of TNT and HMX explosives.

Collection of observations. TABLE II

Comp.	Sugar	PABA	Salic.	Pic.	RDX	TNT/HMX
No.	119	181	135	152	135	197

Records were randomly divided into training and test sets. Then, decision tree algorithm was used to classify observations. Application of decision tree both for feature selection and classification confirmed advantages of this technique for data analysis.

In our experiment we reached classifier yields 3% error on the test set. Table III shows results of classification. The rows represents number of observations of class, and columns give us result of classification. For example, for 39 observations of sugar class, 34 were classified correctly and 5 were recognized as a PABA class of materials.

Result of classifications. TABLE III

	Sugar	PABA	Salic.	Picr.	RDX	TNT/HMX
Sugar	34	5				
PABA		58			1	
Salic.			46			
Picr.		1		50		
RDX					47	
TNT/HMX						64

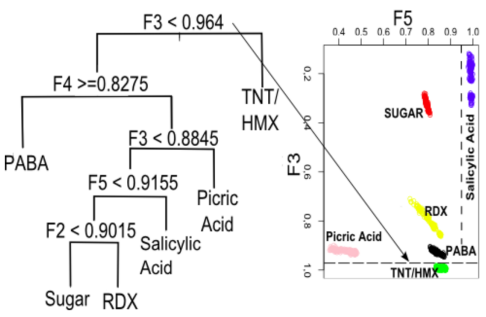

Fig. 7. Clustering observation based on decision trees algorithm.

Another interesting properties were observed. Using two or three classification rules we can make procedure similarity to other discrimination methods like linear discriminant analysis (LDA) or principle component analysis (PCA). A plot of clustering observation is found in Fig. 7. Classification rule $F3 < 0.964$ allowed to separate the TNT/HMX observations, for example.

5. Summary

The samples of RDX, sugar, picric and salicylic acid and PABA were pressed into pellets in pure form, and the pellets were measured in normal-incidence and in 5° reflection geometry using stand-off purged box in TDS system. The spectra were applied to obtain collection of predicted deformed spectra by numerical models of atmosphere. In this article we presented effectiveness of decision tree method for classification THz spectra. The results of our investigation confirmed fitness for purpose of classification THz spectra especially for stand-off detection system.

References

[1] P.F. Tribe, D. Newnham, P. Taday, M. Kemp, *Proc. SPIE* **5354**, 168 (2004).

[2] H. Yang, Z. Xu, J. Zhang, J. Cai, in: *IEEE, Proc. CASoN,* 2010, p. 49.

[3] D. Brigida, X. Zhang, *IEEE Trans. Terahertz Sci. Technol.* **2**, 493 (2012).

[4] R. Ryniec, M. Piszczek, M. Szustakowski, *Acta Phys. Pol. A* **118**, 1235 (2010).

[5] N. Palka, *Acta. Phys. Pol. A* **120**, 715 (2011).

[6] B.D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge 1996.

[7] M. Walesiak, E. Gatnar, *Data Statistical Analysis Using R Program*, PWN, Warszawa 2009 (in Polish).