PCA-K-means Based Clustering Algorithm for High Dimensional and Overlapping Spectra Signals

Nian Zhang, Keenan Leatham
Dept. of Electrical and Computer Eng.
Univ. of the District of Columbia
Washington, D.C. 20008 USA
nzhang@udc.edu, keenan.leatham@udc.edu

Jiang Xiong, Jing Zhong
College of Computer Science and Eng.
Chongqing Three Gorges University
Chongqing, 404000, China
xjcq123@sohu.com, zhongandy@sohu.com

***Abstract*—This paper applied a PCA-K-means method to exploit photo-thermal infrared imaging spectroscopy based trace explosives with overlapping spectral absorption bands. We intend to explore the underlying patterns that affect the clustering performance using top principal components. We also strive to investigate the effectiveness of the clustering algorithm on different analytes and substrates. We reduced the dimensions by applying the principal component analysis (PCA) on the data to transform the original data to the top principal components' feature space. The data were revealed in the feature space and formed into clusters. Then we used the K-means based clustering algorithm to classify them into six classes including RDX, PC, Copper/Steel, TNT, DNT, and PE. After that, we conducted the performance evaluation. We found that the F1 score of the classification of RDX, PC, Copper/Steel, TNT, DNT, and PE is 85%, 39%, 71%, 99%, 92%, and 18%, respectively. The results demonstrated that the proposed algorithm can effectively reduce dimension and accurately determined the classes of those analytes and substrates.**

Keywords—classification; clustering; principal component analysis; k-mean clustering; big data; high dimensional; overlapped spectra

I. INTRODUCTION

Trace analyte detection has become an emergent goal in the fields of military, homeland security, and law enforcement. It provides an early warning of concealed threats and therefore can save people's lives and protect the public facilities. This technology includes remote detection systems capable of detecting explosives and other hazardous materials from a standoff distance. As the demand from military and security markets has increased, it is imperative to develop advanced remote trace detection systems to effectively detect hidden trace explosives in public areas, such as suicide, leave-behind, and vehicle-borne explosives in airports, railway, ship, bus, truck, container, bridge, tunnel, tower, and terminal environments. These remote trace detection systems are also extremely important to identify explosives in the forensic field for crime-scene reconstruction. Further, the remote trace detection systems are also important in agricultural applications, such as monitoring the soil, groundwater, soil gas, surface water, sediment that are suspected of being contaminated.

There is an emerging thrust to replace the traditional handheld explosive trace detectors with standoff sensing systems at a safe distance. Improving survivability and situational awareness has spurred a wide variety of recent development of sensor technology. Most of them are laser-based trace detection systems, such as laser-induced breakdown spectroscopy (LIBS), Raman spectroscopy, laser-induced-fluorescence spectroscopy, and Fourier transform infrared (IR) spectroscopy. Unlike the above traditional detection techniques, the photo-thermal infrared imaging spectroscopy (PT-IRIS) technology can remotely detect trace materials on relevant substrate surfaces from significant standoff distances. In this technique, the surface of interest is illuminated by a light with a specific infrared wavelength and the thermal response of the surface is viewed with an infrared camera [1]. Comparing the thermal image as a function of excitation wavelength of the light to the collection wavelength of surface residues would indicate the presence and location of trace residues. In addition, by changing the excitation wavelength of the light, other trace analytes of interest, such as drugs and chemical agents could also be imaged. Superior to other trace detection techniques, this technology has the potential to generate thermal images of the trace residues and the surface with a spatial resolution of ~1um.

However the ability to detect small amount of residues on large relevant substrate can be very complicated in view of the overlapping optical and thermal spectrum. The key challenge of remote trace detection techniques is to distinguish surface residues from the relevant substrate, such as glasses, paint, and clothes, etc. While substrate materials are chemically different from surface residues, they nonetheless have overlapping IR spectrum. Things become more even worse if the substrate is made of polymeric materials, since such material will absorb the IR spectrum. These real-world challenges add complications to the detection of surface residues.

The advancement in infrared (IR) and Raman spectroscopy has produced numerous massive data, which has generated an urgent need for new spectral analysis techniques. The emerging photo-thermal infrared imaging spectroscopy (PT-IRIS) technique which allows for further increase of the spatial resolution from the current ~10 microns to ~1 micron makes this data analysis demand more critical. This paper will focus on the PT-IRIS data analysis which was used for the application of trace analyte detection. The aim of trace analyte detection is to distinguish illicit surface residues such as explosives from the surface on which they rest. Until now, less effort has been made to develop efficient machine learning techniques to analyze the photo-thermal infrared data.

The rest of the paper is organized as follows. In Section II, the high dimensional and overlapped data set is described. In Section III, the proposed methodology is presented. A combined principal component analysis (PCA)-K-means clustering algorithm is explained. In Section IV, the analysis and results are demonstrated. In Section V, the conclusions are given.

II. Photo-Thermal Infrared Imaging Spectroscopy

This section will use a PCA-K-means method to exploit PT-IRIS based trace explosives with overlapping spectral absorption bands. We intend to explore the underlying patterns that affect the clustering performance using top principal components. We also strive to investigate the effectiveness of the clustering algorithm on different analytes and substrates.

A. Data Set

The advanced photo-thermal infrared imaging spectroscopy (PT-IRIS) technique that can be used for standoff detection application [1]. The two fundamental components are infrared (IR) quantum cascade lasers (QCL) and IR focal plane array detectors. Specifically, IR QCL is used to illuminate the surface residues. If the excitation wavelength of the light is resonant with the collection wavelength of the surface residues, the residues of interest will heat up by (~1ºC). The IR focal plane array detectors are used to provide imaging system.

The temperature increase at each laser pulse, denoted as Tmax is defined as a function of excitation and collection wavelengths. The normalization of Tmax to the average power of the laser pulse will then be used as feature vectors, as shown in Fig. 1. Simulated samples include 5 analytes (TNT, DNT, RDX, Polyethylene, and Polycarbonate) on 4 substrates (Copper, Steel, Polyethylene, and Polycarbonate) using 28 excitation wavelengths (6.0 μm to 6.6 μm and 7.0 μm to 7.7 μm) and 26 collection wavelengths (8.0 μm to 10.5 μm).

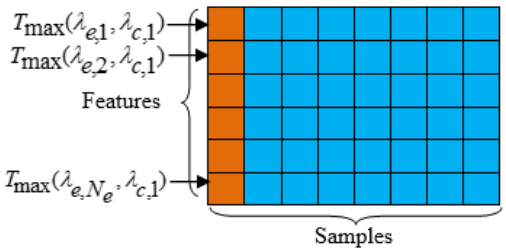

Fig. 1 Data matrix with feature vectors. The *Tmax* values for different excitation and collection wavelength combination were made into feature vectors (columns).

The fundamental spectroscopic characteristics for the PT-IRIS is shown in Fig. 2. It shows the IR absorbance spectra of various materials at different frequency. Peaks in the curves reflect unique "signatures" for each analyte. According to Kirchhoff's Law, the emissivity of a material and its absorptivity are equivalent at thermal equilibrium. Thus, the absorption spectrum can be used to accurately predict its emission spectrum. In another word, if we can determine the most important features (i.e. Tmax values) among the 728 features, which are correlated with the absorption spectrum of a material, we can use the absorption spectrum to predict its emission spectrum. Since the thermal emission from analyte of interest and the surface have different spectral signatures, the unique thermal emission spectrum can ultimately determine the type of this material.

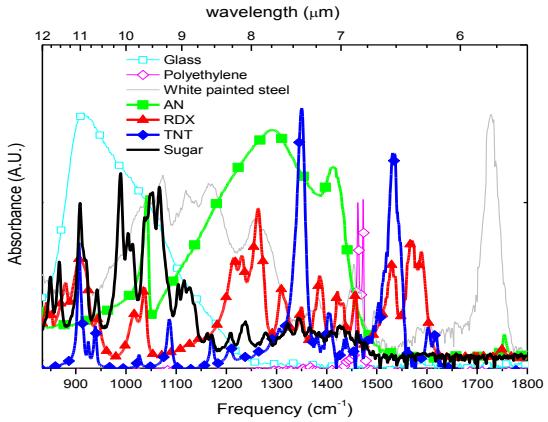

Fig. 2 IR absorbance spectra of glass, polyethylene, white painted steel, ammonium nitrate, RDX, TNT and sucrose (sugar).

Each pixel is a column in this data matrix, which includes 728 features. For each feature, it is a function of Tmax in terms of different combinations of excitation wavelengths and collection wavelengths. With 28 excitation wavelengths and 26 collection wavelengths, they would generate 728 different combined features. Therefore, we may demonstrate the photo-thermal signal matrix for all the 468 samples. This can be seen by display ,the data set in false color plot ,which will show visible or non-visible parts of the electromagnetic spectrum, shown in Fig. 3. From left to right, the particle sizes are 8 μm, 12 μm, 20 μm, 3 μm, 1.5 μm, and 5 μm. Each loop takes 76 columns per particle size. Each column contains 728 features. In Fig. 3, the color is proportional to signal strength, i.e. red represents high, and blue represents low.

For each analyte including TNT, DNT, and RDX, they will be made of all 6 possible particle sizes, and two pixels in the camera frame (i.e. columns) are on the particle, the two pixels will rest on all 4 substrates (i.e. copper, steel, PC, and PE). Thus, there will be a maximum of 48 samples for each analyte. In addition, for each substrate including copper, steel, PC, and PE, they will spread over the 6 particle sizes with two pixels off each particle, and they interact with 5 analytes (TNT, DNT, RDX, PC, and PE), so the maximum of substrate samples is 60.

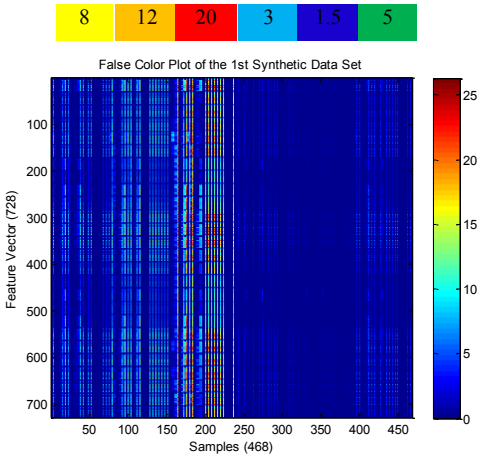

Fig. 3 Six clusters consisting of the four analytes and two substrates were formed using the K-means clustering method.

PC can be used for both analyte and substrate, so the maximum sample amount would be the total of possible analyte and substrate samples, which is therefore 108. PC has a strong signature. PE can also be used for both analyte and substrate. There will be 108 samples at maximum. PE is known as a poor thermal conductor, as a result the temperature increase with illumination is nearly zero. Thus it does not have its own signature.

Since the samples comprise both "on" and "off" particle pixels, some samples may have analyte, some may have substrate, but some complicated and overlapping sample can happen by optically "through" a particle. The mixing IR absorption/emission features reflects the primary challenge to a useful detection technique in the real-world application.

III. Research Methodology

Feature selection is a very important pre-processing technique for large scale pattern recognition problems, especially when the number of available samples is relatively small [2]-[4]. Feature selection is used to find a subset of original features to facilitate optimization, clustering, and classification without a significant loss of accuracy [5][6]. The features contain relevant, irrelevant, and unused information. However, irrelevant and redundant features are useless, which may cause significant degradation in performance due to large search space known as "the curse of dimensionality". By eliminating useless features in the pre-processing stage, feature selection technique could help reduce computational complexity and the effect of curse of dimensionality, and improve the prediction accuracy [7]-[13].

A. Principle Component Analysis (PCA)

Principle component analysis (PCA) is quantitatively rigorous method for achieving dimensional reduction before applying the feature selection methods. It is capable of revealing and identifying patterns in data [14]. The method generates a new variable set, denoted as principal components. Each principal component can be represented as a linear equation of the original variables. Since all the principal components are orthogonal to each other, so there will be no redundant information. Several top ranking principal components are often selected to form a new feature space. The original data will be mapped to this new feature space in the directions of the principal components. Although the PCA can effectively reduce the number of dimensions by selecting the top ranking principle components, PCA method is not able to select a subset of features which are important to distinguish the classes. It only guarantees that when you project each observation on an axis (along a principle component) in a new space, the variance of the new variable is the maximum among all possible choices of that axis. This means that each feature is considered separately, thereby ignoring feature dependencies, which may lead to worse classification performance.

In the application of trace analyte detection where class labels are often unavailable, the feature selection becomes extremely difficult in such unsupervised learning scenario. The feature selection technique is essentially a combinatorial optimization problem which is computationally expensive. The existing and most powerful unsupervised feature selection technique is principle component analysis (PCA). It is often useful to map data onto their principal components rather than on the original x-y axis. In this way the underlying structure in the data can be identified. We applied the PCA technique to the data set to reveal the patterns in data, as well as reduce the dimension of feature vectors (i.e. vectors containing the principle components). First we deconstruct the set into eigenvectors and eigenvalues. An eigenvector is a direction, and an eigenvalue is a number, telling you how much variance there is in the data in that direction. The amount of eigenvectors/values is the same as the dimensions that the data set has. The reason for this is that eigenvectors convert the data into a new set of dimensions, and the number of dimensions have to be equal to the original amount of dimensions. It is worthwhile to investigate the PCA algorithm because it allows us to exploit the correlation of most significant eigenvectors and analyte types.

B. PCA-K-means Clustering Algorithm

PCA algorithm will be applying to reduce the dimensions, and then the k-means clustering algorithm will be applied.

Steps	**K-means Clustering Algorithm**
1	k initial "means" (k is an estimated value) are randomly generated.
2	k clusters are formed by assigning an observation to its nearest mean.
3	The centroids of k clusters become the new mean.
4	Repeat steps 2 and 3 until convergence.

IV. ANALYSIS AND RESULTS

A. PCA-K-means Clustering Results

We presented the principal component analysis (PCA) results using a combination of top principal components (PCs), i.e. PC1 and PC2. We displayed the data in PC1-PC2 axes, as shown in Fig. 4.

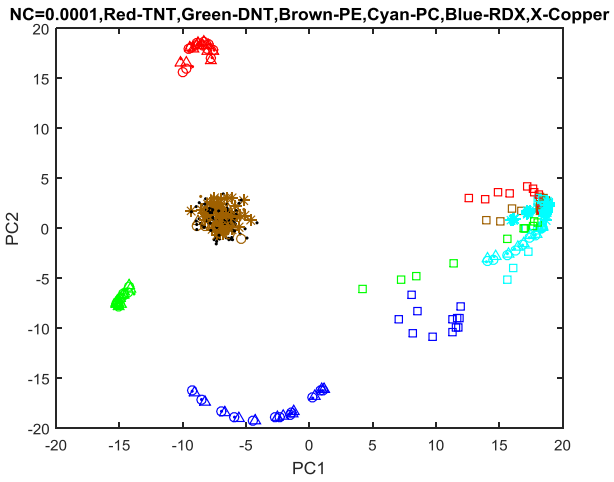

Fig. 4 All the data projected onto PC1-PC2 space after using the PCA method.

Then we applied the K-means clustering algorithm to make them into 6 clusters. The number of analytes and substrates in each of the six classes are shown in Table I. Letter C represents the cluster. The rows represent TNT, DNT, PE, PC, RDX, and Copper, respectively. The columns represent Cluster 1, Cluster 2, Cluster 3, Cluster 4, Cluster 5, and Cluster 6.

Table I. Number of Samples in Classifiers after Using the PCA-K-means Clustering Algorithm

	C1	C2	C3	C4	C5	C6
TNT	0	10	0	36	0	0
DNT	0	6	0	0	36	0
PE	0	12	96	0	0	4
PC	0	104	0	0	0	0
RDX	34	0	0	0	0	12
Copper	0	0	118	0	0	0

By counting the largest number of analytes/substrate in each class, we can find the analyte/substrate dominating that class. For each cluster, the majority analytes will determine the class that this cluster belongs to. Therefore, we find Cluster 1 represents RDX, Cluster 2 represents PC, Cluster 3 represents Copper, Cluster 4 represents TNT, Cluster 5 represents DNT, and Cluster 6 represents PE. These labels are shown in the tables in the following Clustering Analysis section.

The classes are demonstrated in PC1-PC2 axes, as shown in Fig. 5. To enhance the visualization efficiency, we adopt the same color code as the PCA algorithm. Red dots represent TNT, Green dots represent DNT, Yellow dots represents PE, Cyan dots represent PC, Blue dots represents RDX, and Black dots represent Copper/Steel. The centroids of the classes are indicated by black crosses. Thus, it is easier to compare Fig. 4 and Fig. 5.

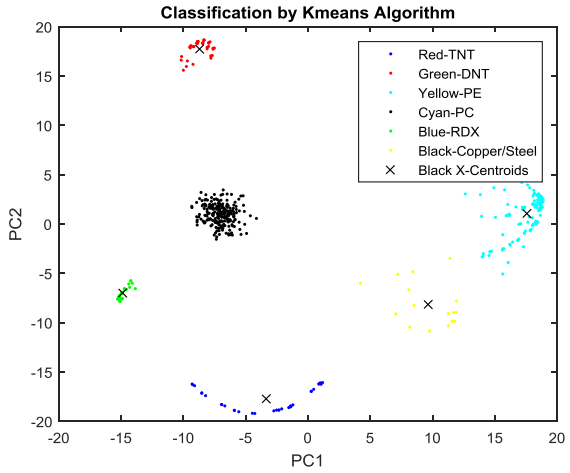

Fig. 5 Six clusters consisting of the four analytes and two substrates were formed using the K-means clustering method.

B. Clustering Analysis

We investigate popular clustering performance evaluation which consists of probability of detection (POD), false alarm rate (FAR), accuracy, precision, recall, and F1 score for each residue of interest. F1 score is usually more useful than accuracy, especially if there exists an uneven class distribution. The procedure involves the determination of true positive (TP), false negative (FN), false positive (FP), and true negative (TN) of residues of interest. In the

pre-processing stage, based on Table I, six tables are generated for the six analytes in Table II-VII. Different colors are chosen for these metrics, as shown below.

Table II. Labeling for TNT

Analytes	Class 1 (RDX)	Class 2 (PC)	Class 3 (Copper)	**Class 4 (TNT)**	Class 5 (DNT)	Class 6 (PE)
TNT	0	10	0	36	0	0
DNT	0	6	0	0	36	0
PE	0	12	96	0	0	4
PC	0	104	0	0	0	0
RDX	34	0	0	0	0	12
Copper	0	0	118	0	0	0

Table III. Labeling for DNT

Analytes	Class 1 (RDX)	Class 2 (PC)	Class 3 (Copper)	Class 4 (TNT)	**Class 5 (DNT)**	Class 6 (PE)
TNT	0	10	0	36	0	0
DNT	0	6	0	0	36	0
PE	0	12	96	0	0	4
PC	0	104	0	0	0	0
RDX	34	0	0	0	0	12
Copper	0	0	118	0	0	0

Table IV. Labeling for PE

Analytes	Class 1 (RDX)	Class 2 (PC)	Class 3 (Copper)	Class 4 (TNT)	Class 5 (DNT)	**Class 6 (PE)**
TNT	0	10	0	36	0	0
DNT	0	6	0	0	36	0
PE	0	12	96	0	0	4
PC	0	104	0	0	0	0
RDX	34	0	0	0	0	12
Copper	0	0	118	0	0	0

Table V. Labeling for PC

Analyes	Class 1 (RDX)	**Class 2 (PC)**	Class 3 (Copper)	Class 4 (TNT)	Class 5 (DNT)	Class 6 (PE)
TNT	0	10	0	36	0	0
DNT	0	6	0	0	36	0
PE	0	12	96	0	0	4
PC	0	104	0	0	0	0
RDX	34	0	0	0	0	12
Copper	0	0	118	0	0	0

Table VI. Labeling for RDX

Analytes	**Class 1 (RDX)**	Class 2 (PC)	Class 3 (Copper)	Class 4 (TNT)	Class 5 (DNT)	Class 6 (PE)
TNT	0	10	0	36	0	0
DNT	0	6	0	0	36	0
PE	0	12	96	0	0	4
PC	0	104	0	0	0	0
RDX	34	0	0	0	0	12
Copper	0	0	118	0	0	0

Table VII. Labeling for Copper/Steel

Analytes	Class 1 (RDX)	Class 2 (PC)	**Class 3 (Copper)**	Class 4 (TNT)	Class 5 (DNT)	Class 6 (PE)
TNT	0	10	0	36	0	0
DNT	0	6	0	0	36	0
PE	0	12	96	0	0	4
PC	0	104	0	0	0	0
RDX	34	0	0	0	0	12
Copper	0	0	118	0	0	0

C. Clustering Performance Evaluation

We then conducted the clustering performance evaluation by calculating the evaluation metrics, including the probability of detection (POD), false alarm rate (FAR), accuracy, precision, recall, and F1 score (the higher the better) for each residue of interest. The above evaluation metrics are defined as follows [9]:

Probability of Detection: POD = TP/(TP+FN)
False Alarm Rate: FAR = FP/(FP+TN)
Precision: P = TP/(TP+FP)
Recall: R = TP/(TP+FN)
Accuracy: Accuracy = (TP+TN)/(TP+FP+FN+TN)
F1 Score: F1 Score = 2 * Precision * Recall/(Precision + Recall)

Regarding the precision measure, it indicates how often an instance was predicted as positive that is actually positive. On the other hand, a recall measures how often a positive class instance was predicted as a positive class instance by the classifier. In imbalanced learning, the goal is to improve recall without a significant loss of precision. However, it is extremely challenging to accomplish this goal, since in order to increase the TP for the minority class, the number of FP is also increased, which will result in a reduced precision.

The k-means clustering algorithm on the data on PC1 and PC2 was performed. The clustering performance including the probability of detection (POD), false alarm rate (FAR), accuracy, precision, recall and F1 score is conducted. The performance results are shown in Table VIII. The six clusters were determined by the majority analyte type in each cluster.

Accuracy can be significantly affected by the number of true negatives which in the application of trace analyte detection, are not as critical indicators as false negative and false positive. Therefore, F1 score is usually a better measure to evaluate if we need to seek a balance between precision and recall and the data has an uneven class distribution.

Virtually there is no signal from polyethylene at any excitation wavelength or collection wavelength. Compared to polyethylene (PE), copper, and steel, only polycarbonate (PC)

has its own "spectrum". Spectral mixing problem becomes worst when the trace analytes rest on such active substrate. It will result in poor F1 score.

TABLE VIII. PCA-K-Means Clustering Performance on PC1 and PC2

	Class1 (RDX)	Class2 (PC)	Class3 (Cop)	Class4 (TNT)	Class5 (DNT)	Class6 (PE)
POD	74%	100%	100%	78%	86%	10%
FAR	7%	8%	27%	0	0	1%
Accuracy	97%	94%	79%	98%	98%	76%
Precision	100%	24%	55%	100%	100%	75%
Recall	74%	100%	100%	98%	86%	10%
F1 Score	85%	39%	71%	99%	92%	18%

V. CONCLUSION

This paper applied a PCA-K-means method to exploit PT-IRIS based trace explosives with overlapping spectral absorption bands. We intend to explore the underlying patterns that affect the clustering performance using top principal components. We also strive to investigate the effectiveness of the clustering algorithm on different analytes and substrates. The principal component analysis (PCA) was used to reduce the dimension of data space to the top principal components feature (PC1-PC2) space, and thus the most prominent features or patterns were revealed. Then we used the K-mean clustering algorithm to classify them into four analytes and two substrates. We used the performance evaluation matrices to measure the accuracy of classification. The experimental results demonstrated that the combination of the principal component analysis and K-means clustering algorithm are efficient for achieving dimensional reduction and clustering on highly overlapped photo-thermal infrared imaging data. The F1 score of the classification of RDX, PC, Copper, TNT, DNT, and PE is 85%, 39%, 71%, 99%, 92%, and 18%, respectively.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation (NSF) grants: HRD #1505509, HRD #1533479, and DUE #1654474.

REFERENCES

[1] C. Kendziora, R. Furstenberg, M. Papantonakis, V. Nguyen, J.Stepnowski, and R. McGill, "Advances in standoff detection of trace explosives by infrared photo-thermal imaging," *Proc. SPIE 7664, Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XV*, 76641J, 2010.

[2] L. Zhang, Q. Zhang, B. Du, X. Huang, Y. Y. Tang, and D. Tao, "Simultaneous spectral-spatial feature selection and extraction for hyperspectral images," *IEEE Transactions on Cybernetics*, vol. 48, no. 1, pp. 16-28, 2018.

[3] M. Liu, C. Xu, Y. Luo, C. Xu, Y. Wen, and D. Tao, "Cost-sensitive feature selection by optimizing f-measures," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1323-1335, 2018.

[4] X. Wen, L. Shao, W. Fang, and Y. Xue, "Efficient feature selection and classification for vehicle detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 3, pp. 508-517, 2015.

[5] M. Dash and H. Liu, "Feature selection for classification," Intell. Data Anal., vol. 1, no. 1-4, pp. 131-156, 1997.

[6] A. Unler and A. Murat, "A discrete particle swarm optimization method for feature selection in binary classification problems," Eur. J. Oper. Res., vol. 206, no. 3, pp. 528–539, Nov. 2010.

[7] N. Zhang and K. Leatham, "Feature selection based on SVM in photo-thermal infrared (IR) imaging spectroscopy classification with limited training samples", *WSEAS Transactions on Signal Processing*, ISSN / E-ISSN: 1790-5052 / 2224-3488, vol. 13, Art. #33, pp. 285-292, 2017.

[8] N. Zhang, J. Xiong, J. Zhong, and K. Leatham, "Gaussian process regression method for classification for high-dimensional data with limited samples", *The 8th International Conference on Information Science and Technology (ICIST 2018)*, Cordoba, Granada, and Seville, Spain, June 30-July 6, 2018.

[9] N. Zhang, J. Xiong, J. Zhong, and L. A. Thompson, "Feature selection method using BPSO-EA with ENN classifier", *The 8th International Conference on Information Science and Technology (ICIST 2018)*, Cordoba, Granada, and Seville, Spain, June 30-July 6, 2018.

[10] N. Zhang and L. A. Thompson, "An intelligent clustering algorithm for high dimensional and highly overlapped photo-thermal infrared imaging data," *Fall 2016 ASEE Mid-Atlantic Regional Conference*, Hofstra University, Hempstead, NY, October 21-22, 2016.

[11] N. Zhang, "Cost-sensitive spectral clustering for photo-thermal infrared imaging data," *2016 Sixth International Conference on Information Science and Technology (ICIST)*, Dalian, pp. 358 – 361, May 6-8, China, 2016.

[12] J. F. Ramirez Rochac and N. Zhang, "Reference clusters based feature extraction approach for mixed spectral signatures with dimensionality disparity," *10th Annual IEEE International Systems Conference (IEEE SysCon 2016)*, Orlando, Florida, pp. 1 – 5, April 18-21, 2016.

[13] J. F. Ramirez Rochac and N. Zhang, "Feature extraction in hyperspectral imaging using adaptive feature selection approach," *The Eighth International Conference on Advanced Computational Intelligence (ICACI2016)*, Chiang Mai, Thailand, pp. 36-40, 2016.

[14] L. I. Smith, A Tutorial on Principal Components Analysis, 2002.