UACEE International Journal of Advances in Computer Science and its Applications – IJCSIA
Volume 3 : Issue 2 [ISSN 2250 – 3765]
Publication Date : 05 June 2013

SAW Sensor Array Data Fusion for Chemical Class Recognition of Volatile Organic Compounds

Sunil K. Jha, Kenshi Hayashi
Department of Electronics, Graduate School of Information Science and Electrical Engineering
Kyushu University, 744 Motooka, Fukuoka 819-0395, JAPAN
drsuniljha@o.ed.kyushu-u.ac.jp, hayashi@ed.kyushu-u.ac.jp

Abstract— **Present study deals the development of data fusion based artificial intelligence unit for the chemical sensor array based electronic nose (E-Nose) system. We focus particularly on feature level fusion of model surface acoustic wave (SAW) sensor array response for chemical class identification of volatile organic compounds (VOCs). Three methods are used for feature extraction namely: principal component analysis (PCA); independent component analysis (ICA) and kernel principal component analysis (KPCA). Fused features are generated with three unsupervised fusion schemes and validated in combination with support vector machine (SVM) classifier. Study is concluded by the analysis of 12 model SAW sensor array data sets. It suggests that amongst the three feature fusion schemes; feature fusion by summation result highest class recognition rate of VOCs.**

***Keywords*—data fusion, saw sensor, electronic nose, chemical class recognition**

I. Introduction

A set of chemical sensors of varied selectivity in combination with the artificial intelligence methods is popularly known as electronic nose (E-Nose). It is used for the recognition of volatile organic compounds (VOCs) in monitoring of food quality, environment, health, and safety to security applications [1]. Surface acoustic wave (SAW) oscillator coated with chemoselective polymer is a well recognised chemical sensor used in E-Nose system [2]. Most of the limitations of present E-Nose system can be minimized by the optimization of sensor array and artificial intelligence system (sensor array information processing unit for chemical vapor class recognition or concentration estimation). The development of an efficient and reliable artificial intelligence unit is the most limiting aspects of E-Nose [1]. With the increase in the complexity of the sensing environment information collected by using a single pattern recognition method at each steps of artificial intelligence unit may not be sufficient and results in low recognition efficiency of E-Nose. Data fusion approach overcomes this limitation by integrating the information extracted by various pattern recognition methods [3−5]. The architecture and procedure selection for data fusion strategy are the domain specific. It is also an open research issue in present E-Nose system [3].

Three major structures of data fusion are reported in literature namely; pre-processing, feature and decision level fusion [3]. In data level fusion raw data set is preprocessed using different preprocessing methods and then fused into a new single raw data set. In feature level fusion numerous feature extraction methods are used to generate the diverse feature vectors in different feature spaces, thereafter fused together to obtain noble feature vectors. In decision level fusion data set is processed with independent classifiers after that a common decision is made by fusion of decision of each of the individual classifiers. Initially data fusion methods have been developed for the military applications like target and threat recognition, remote sensing, battlefield surveillance etc. [6] However at the present these methods have also been widely used in image processing, face recognition, speech processing, video classification and retrieval, gene detection and E-Nose etc [7−12].

The application of data fusion in E-Nose domain is reported in some of the earlier studies [8−11]. Dutta et al. [8] have used data level fusion in tin oxide sensor array. In another study Natale et al. [9] have presented data level fusion of two varieties of sensors (tin oxide and QCM). Li et al. [10] have reported feature and decision level fusion; a similar approach is presented by L. Rong et al. for wine classification in [11].

The intention of present study is to achieve the optimum performance E-Nose system intelligence by fusing the information from multiple feature extraction methods using simpler approaches. Feature level fusion reduces the commensurate requirement of data level fusion. It also results additional information gain as compared to decision level fusion. This study focuses on feature level data fusion for chemical class recognition of VOCs by analyzing model SAW sensor array response. The feature vectors are generated with three unsupervised feature extraction methods including PCA, KPCA and ICA. These feature vectors are concatenated for fusion by three simple approaches namely: fusion by features summation; features multiplication and features combination. Efficacy of fused feature vectors is validated by using support vector machine (SVM) classifier for class recognition of VOCs. Study is concluded by analyzing 12 sets of SAW sensor array response generated by using SAW sensor model simulation. Rest part of the paper is organized as follows. Detail description of data sets, preprocessing methods, feature extraction methods, feature fusion schemes and classification method are presented in section *II*. Section *III* covers the analysis outcomes of data sets. Section *IV* presents the discussion of research findings and finally the conclusion of study is summarized in section *V*.

II. Data and Processing Methods

A. Data Sets

Data sets are generated by simulation using the model of SAW sensor; combined with different intensity of additive noise and outliers. Each of the data set is based on response of 11-element SAW sensor array (functionalized with different polymers) for 180 chemical vapor samples belonging to six chemical classes. The chemical vapor sample belongs to chemical classes: trinitrotoluene (TNT); dinitrotoluene (DNT); dimethyl methylphosphonate (DMMP); water; toluene and benzene. Sensor array response is computed at 30 different vapor concentration (varies in between ppth (parts per thousands) − ppt (parts per trillion)) of each chemicals. A summary of data sets is presented in Table 1. The basic distinction amongst the data sets is in the value of noise and outliers incorporated in it. For instance the data set-IV contains additive Gaussian noise with mean value 0 Hz and standard deviation 200 Hz , whereas the data set-V includes Gaussian noise with random mean value in between +50 Hz to -50 Hz and standard deviation 50 Hz .More details about the sensor array response generation by simulation can be seen in our earlier study [13].

TABLE I. SAW MODEL SIMULATED DATA SETS USED IN ANALYSIS

Data sets	Noise level Data sets (Mean value and Standard deviation in Hz)	Outlier level in Data sets (Mean value in Hz, Probability of Outlier Addition in %)
I	0, 50	0, 0
II	0, 100	0, 0
III	0, 150	0, 0
IV	0, 200	0, 0
V	Between +50 to –50, 50	0, 0
VI	Between +50 to –50, 100	0, 0
VII	Between +50 to –50, 150	0, 0
VIII	Between +50 to –50, 200	0, 0
IX	0, 10	Between +50 to –50, 75
X	0, 10	Between +50 to –50, 80
XI	0, 10	Between +50 to –50, 85
XII	0, 10	Between +50 to –50, 85

B. Data Preprocessing

SAW sensor array response is defined as Δf_{ij} (change in frequency of *j*-th sensor due to exposure of *i*-th vapor sample). A schematic diagram for processing the sensor array response is given in Fig.1. Preprocessing is completed in three steps. This includes first normalization with respect to frequency shift Δf_p^j due to polymer coating, second logarithmic scaling as suggested in our earlier studies [14, 15] and finally dimension autoscaling described in [16].

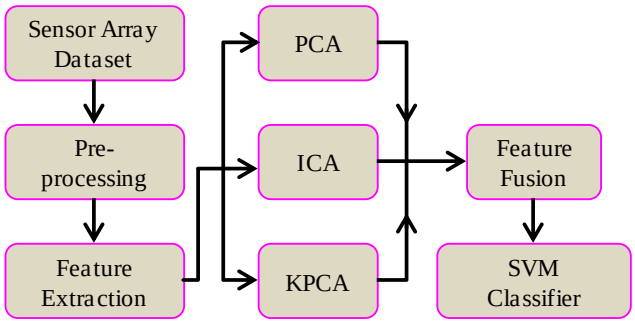

Figure 1. *Flow chart of sensor array response processing.*

C. Feature Extraction

Three feature extraction methods: principal component analysis (PCA); independent component analysis (ICA), and kernel principal component analysis (KPCA) have been used in feature extraction analysis. PCA is a linear unsupervised feature extraction method. It transforms correlated sensor array response in measurement space into uncorrelated principal component (PC) space. KPCA is a nonlinear unsupervised feature extraction method. It maps sensor array response from measurement space into high dimensional space (H-space) by using some nonlinear kernel function. Thereafter linear PCA is implemented on covariance matrix in H-space for feature vector generation. ICA is a linear unsupervised feature extraction method. It finds out orthogonal directions for feature extraction along which the projected sensors response has minimum correlation as well as the statistical dependency. ICA is reported for sensor array signal preprocessing in a most recent study [17]. The detail mathematical description of three discussed feature extraction methods can be found in [18, 19]. '*Stats*' package [20], '*Kernlab*' package [21] and '*Fast ICA*' package [22] available in open source statistical computing language 'R' are used for the implementation of PCA, KPCA and ICA methods respectively.

D. Feature Fusion

Let $R=(R_1,R_2,R_3,.....R_n)$ be the response vector of any chemical vapor from n-element SAW sensor array. Each of the response vectors are projected into PCA, KPCA and ICA spaces. After feature extraction by the three methods $R^{PCA}=\left(R_1{}^{PC1},R_2{}^{PC2},R_3{}^{PC3},...R_n{}^{PCn}\right)$ $R^{KPCA}=\left(R_1{}^{KPC1},R_2{}^{KPC2},R_3{}^{KPC3},....R_n{}^{KPCn}\right)$ and $R^{ICA}=\left(R_1{}^{IC1},R_2{}^{IC2},R_3{}^{IC3},.....R_n{}^{ICn}\right)$ be the corresponding feature vectors in three spaces respectively.

For feature vector fusion three fusion rules have been implemented inspired from the study presented in [12, 23] as:

a. Feature vector summation

$$R_S{}^{fused}=\left(\left(R_1{}^{PC1}+R_1{}^{KPC1}+R_1{}^{IC1}\right),\left(R_2{}^{PC2}+R_2{}^{KPC2}+R_2{}^{IC2}\right),\left(R_3{}^{PC3}+R_3{}^{KPC3}+R_3{}^{IC3}\right).....\left(R_n{}^{PCn}+R_n{}^{KPCn}+R_n{}^{ICn}\right)\right)$$

b. Feature vector multiplication

$$R_M{}^{fused} = \left(\left(R_1{}^{PC1} \times R_1{}^{KPC1} \times R_1{}^{IC1} \right), \left(R_2{}^{PC2} \times R_2{}^{KPC2} \times R_2{}^{IC2} \right), \left(R_3{}^{PC3} \times R_3{}^{KPC3} \times R_3{}^{IC3} \right) \ldots \ldots \left(R_n{}^{PCn} \times R_n{}^{KPCn} \times R_n{}^{ICn} \right) \right)$$

c. Feature vector combination

$$R_C{}^{fused} = \left(R_1{}^{PC1}, R_1{}^{KPC1}, R_1{}^{IC1} \right)$$

We hardly found these fusion rules in E-Nose data processing. The dimension of each of the data set is 11. After feature extraction we selected only first three dimensions. Thus feature vectors generated by three above mentioned fusion rules have also only three dimensions. It will also help in reducing the computation time of analysis as well as in comparative analysis amongst the three fusion schemes.

E. Classification

In present study Support vector machine (SVM) method is employed by using the 'e1071' package [24] in 'R' for class identification of chemical vapor by using the chemical feature vectors as input. This method is introduced by Vapnik [25] and summarized in review [26]. In binary class recognition problem, the method builds an optimal separating hyperplane using the training feature vectors. The hyperplane maximizes the interclass margin by using the quadratic programming (QP) optimization technique. The training feature vector close to hyperplane are used to measure the interclass margin and known as support vector. For multiclass identification, training feature vectors are divided into combinations of several binary classes and SVM model is trained with each. In validation the test feature vectors are classified with binary class trained SVM models. The final decision for the class of an unknown feature sample is made on the basis of majority voting of the binary class models.

III. Analysis Outcomes

Each of the data sets is processed according to the analysis flow chart shown in Fig.1. The data sets are preprocessed using the methods discussed in section II *B*. Next, feature vectors are generated using the three unsupervised feature extraction methods: PCA; ICA and KPCA analyzing each of the data set independently. After feature extraction with PCA, ICA and KPCA we have selected only three features. That is the dimensionality of data set is reduced from 11→3 in feature space. In fusion only 3–dimensional feature vectors from each of the three features space is used. The fused feature vectors are computed by experimenting with three suggested fusion schemes discussed in II *D*. This result 3−dimensional fused feature vectors. To check the chemical class recognition ability of single as well as fused feature vectors, SVM classifier is employed. Each of the data sets whether in single and fused feature space is divided first into training and test sets. 2/3rd of feature vectors (120 samples) are used in training of SVM classifier and remaining 1/3rd (60 samples) are used for its validation. Correct class recognition rate by the SVM classifier for all the 12 data sets in both the single and fused feature space is summarized in Table 2.

IV. Discussion

Two main varieties of reported data fusion rules are: 1) unsupervised rules (sum, product, minimum and maximum) and ii) supervised rules (SVM, bagging and boosting) [32]. Present study implements only unsupervised fusion rules since they can be easily executed without training. The summary of SVM classification results for all the 12 data set are presented in Table 2. Each digit in Table 2 represents the correct class recognition rate in validation phase of the SVM classifier using the features in single and fused feature spaces.

TABLE II. SVM CLASSIFICATION RESULTS IN SINGLE AND FUSED FEATURE SPACES.

Data Sets	True class recognition rate in % by SVM classifier using					
	Single features by			Fused features By		
	KPCA	ICA	PCA	Multiplication	Combination	Addition
I	58.3	78.3	91.7	53.3	91.7	93.3
II	48.3	58.3	88.3	53.3	93.3	93.3
III	53.3	71.7	86.7	43.3	70.0	98.3
IV	46.7	66.7	81.7	45.0	90.0	91.7
V	53.3	63.3	81.7	35.0	75.0	86.7
VI	53.3	46.7	86.7	45.0	63.3	81.7
VII	55.0	58.3	86.7	33.3	76.7	86.7
VIII	33.3	55.7	51.7	36.7	66.7	78.3
IX	53.3	71.7	88.3	60.0	95.0	91.7
X	63.3	66.7	88.3	58.3	76.7	95.0
XI	53.3	21.7	43.3	38.3	63.3	61.7
XII	48.3	25.0	56.7	38.3	56.7	71.7

It is evident from the Table 2 that the class recognition capability of feature vector both in single and fused feature space decreases from data sets-I–XII. This is reasonable since as going from the data sets-I–XII the level of noise and outliers incorporated in data sets increases. Also amongst the three kinds of feature vectors in single feature space, the performance of PCA features is better as compare to ICA and KPCA features. Amongst the three fused features the additive features result average correct classification rate more than the 90% except for the data sets XI and XII. Since these two data sets have additional amount of noise and outlier. The feature vectors generated by multiplication have poor class recognition capability with an average class recognition rate approximately 50%. Its performance looks equivalent to the KPCA feature vectors in single feature space. The feature generated by the combination performs equally well as additive features in case of less noisy data sets but as the noise level in data sets increases its performance decreases (see the results of data sets-I–IV and data sets-V–XII in Table 2).

Again if we compare the class identification capability of PCA in single feature space and additive feature vectors in fused feature space; additive feature vectors looks to have

better recognition capability. Finally it can be concluded from Table 2 that the additive features generated by summation have the best class identification capability in combination with SVM classifier in fused space.

This is due to the additive Gaussian nature of noise included in each of the data set and its elimination by feature extraction method and feature fusion. The effect of noise is suppressed by each of the feature extraction methods independently when dimensionality is reduced. The effect of noise is further reduced in the features generated by the summation (since the residual noise is just added) as compare to the features generated by the multiplication (the residual noise is multiplied). This may be the possible explanation for better performance of additive features compare to the multiplicative features. Also the features generated by combination have the higher noise level as compare to the additive features and lower noise level as compare to the multiplicative features. Since in combination the residual noise is get combined with each of the feature dimensions. Thus the features generated by the combination have better chemical class recognition efficiency than the features generated by the multiplication and poor efficiency than the features generated by summation.

v. Conclusion

The present study explores elementary feature fusion approaches for chemical vapor class recognition based on response analysis of model SAW sensor array in combination with SVM classifier. Analysis outcomes represents that the chemical vapor recognition efficiency of SAW sensor array based E-Nose can be enhanced by using the fused feature vectors. Feature fusion by addition is found to be an effective fusion approach. It groups the discriminating information of chemical vapor from different feature spaces by eliminating the noise.

Acknowledgment

This research is funded by JSPS (Japan Society for the promotion of Science) related to the JSPS Grant-in-Aid for foreign researcher Post Doctoral Fellowship (P12367). The author S.K.J gratefully acknowledges Director, A.K.G. Engineering College, Ghaziabad, India, Prof. P.K. Sharda for their consent and motivation to pursue this study, all my colleagues and Mrs. Anju Sunil Jha for their cooperation and support .

References

[1] J.W. Gardener and P.N. Bartlett, Electronic Noses Principles and Application, New York, Oxford University Press, 1999.

[2] K. Arshak, E. Moore, G.M. Lyons, J. Harris, and S. Clifford, "A review of gas sensors employed in electronic nose applications", Sensor Review, vol. 24, pp. 181–198, 2004.

[3] I. Bloch, "Information combination operators for data fusion: a comparative review with classification", IEEE Trans. Syst. Man Cyber. Vol. 26, pp. 52–67,1996.

[4] P.K. Varshney, Distributed Detection and Data Fusion, Springer-Verlag, New York, 1997.

[5] D.L. Hall and J. Llinas, "An introduction to multisensor data fusion", IEEE Proc., vol. 85, pp. 6–23, 1997.

[6] D. Smith and S. Singh, "Approaches to Multisensor Data Fusion in Target Tracking: A Survey", IEEE Trans. Know. Data Eng., vol. 18, pp. 1696−1710, 2006.

[7] S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Fusion of Face and Speech Data for Person Identity Verification", IEEE Trans. Neural Networks, vol. 10, pp. 1065−1074, 1999.

[8] R. Dutta, E.L. Hines, J.W. Gardner, D.D. Udrea, and P. Boilot, "Non-destructive egg freshness determination: an electronic nose based approach", Meas. Sci. Technol., vol. 14, pp. 190–198, 2003.

[9] C.D. Natale, A. Macagnano, S. Nardis, R. Paolesse, C. Falconi, E. Proietti, P. Siciliano, R. Rella, A. Taurino, and A.D' Amico, "Comparison and integration of arrays of quartz resonators and metal oxide semiconductors chemiresistors in the quality evaluation of olive oils", Sens. Act. B, vol. 78, pp. 303–309, 2001.

[10] C. Li, P. Heinemann, and R. Sherry, "Neural network and Bayesian network fusion models to fuse electronic nose and surface acoustic wave sensor data for apple defect detection", Sens. Act. B, vol. 125, pp. 301–310, 2007.

[11] L. Rong, W. Ping, and H. Wenlei, "A novel method for wine analysis based on sensor fusion technique", Sens. Act. B, vol. 66, pp. 246–250, 2000.

[12] G. Marcialis and F. Roli, "Decision level fusion of PCA and LDA based face recognition algorithms", International Journal of Image and Graphics, vol. 6, pp. 239–311, 2006.

[13] S.K. Jha and R.D.S. Yadava, "Development of surface acoustic wave electronic nose using pattern recognition system", Defence Science Journal, vol. 60, pp. 364–376, 2010.

[14] R.D.S. Yadava and R. Chaudhary, "Solvation transduction and independent component analysis for pattern recognition in SAW electronic nose", Sens. Act. B, vol. 113, pp.1–21, 2006.

[15] S.K. Jha and R.D.S. Yadava, "Preprocessing of SAW sensor array data and pattern recognition", IEEE Sensors Journal, vol. 9, pp. 1202–1208, 2009.

[16] R.G. Osuna and H.T. Nagle, "Method for evaluating data-preprocessing techniques for odor classification with an array of gas sensor", IEEE Trans. System Man Cybernetics: B, vol. 29, pp. 626–632, 1999.

[17] M. Imahashi and K. Hayashi, "Odor clustering and discrimination using an odor separating system", Sens. Act. B, vol. 166-167, pp. 685−694, 2012.

[18] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, USA, 2006.

[19] A. Hyvärinen, J. Karhunen, and E. Oja, Independent Component Analysis, John Wiley & Sons, Canada, 2001.

[20] R Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051- 07-0, 2008. http://www.R-project.org.

[21] A. Karatzoglou, A.Smola, K. Hornik, A. Zeileis, Kernlab - An S4 Package for Kernel Methods in R, Journal of Statistical Software 11, (2004) 1–20. http://www.jstatsoft.org/v11/i09 .

[22] J.L. Marchini, C. Heaton, B.D Ripley, fastICA: FastICA Algorithms to perform ICA and Projection Pursuit. R package version 1.1-9, 2007.

[23] G. Fumera and F. Roli, "Analysis of error-reject trade-off in linearly combined multiple classifiers", Pattern Recognition, pp. 1245–1265, 2004.

[24] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, A. Weingessel, e1071: Misc Functions of the Department of Statistics (e1071), TU Wien, R package version 1.5-18. 2008.

[25] V. Vapnik, The Nature of Statistical Learning Theory, Springer Verlag, New York, 1995.

[26] J.C. Christopher and A. Burges, "Tutorial on support vector machines for pattern recognition", Data Mining and Knowledge Discovery, vol. 2 pp. 121–167, 1998.