Source Classification of Indoor Air Pollutants using Principal Component Analysis for Smart Home Monitoring Applications

Y.-T. Chen[1] and S. Shrestha[2*]
[1]*Department of Electrical and Computer Engineering, Miami University,* Oxford, OH, USA 45056
2Department of Engineering Science, Sonoma State University, Rohnert Park, CA, USA 94928

[1]cheny97@miamioh.edu, [2]sudhir.shrestha@sonoma.edu (*Corresponding Author)

***Abstract*—Indoor air pollution has much greater impact on our health than we perceive. Presence of particulates and volatile organic compounds (VOCs) in indoor air are in much higher concentrations than outdoor air. These particulates and VOCs are known to cause numerous health problems to millions of people every year. This paper presents a solution for passive and continuous monitoring of harmful VOCs using a sensor array. We tested common household products for VOCs emissions. The items were tested in a controlled laboratory setting which simulated an indoor environment. In the laboratory, the developed system was able to detect the presence of the harmful VOCs and classify the sources of those VOCs. Principal component analysis (PCA) was used for identification and classification. The presented system aims to assist the users to monitor the presence of harmful VOCs and inform about their possible sources. How we designed the system and the test results are presented and discussed.**

Keywords—Smart sensing system, indoor air quality monitoring, volatile organic compounds, sensor array, principle component analysis

I. INTRODUCTION

The indoor air quality monitoring has been widely studied for the presence of volatile organic compounds and for toxic and unhygienic conditions [1]. Studies have shown that the concentrations of some compounds and particles in indoor air are two to five times higher than outdoor air [2]. Studies have also shown that, every year, there are 1.5 to 2 million deaths worldwide that are linked to indoor air pollution [3]. According to one study, in the United States, people spend majority of their time inside buildings or vehicles [4]. As we spend more time indoors than outdoors, indoor air quality plays a vital role in our health and well-being. VOCs, emitted by many common household items, are a major source of indoor air pollution. Some of those VOCs are toxic to humans, and are known to have harmful effects to our health [5, 6]. A study conducted by the U.S. Environmental Protection Agency in 1987 indicated that indoor air pollution is the fourth top carcinogen. Some other common adverse effects that people experience when they live in poor indoor air quality include, allergic reactions or long-term damage to kidneys, liver, and nervous system [10]. With increasing regular usage of household products such as pest and insect repellents, cleaning products, and odor neutralizers, the risk of continued exposure to toxic compounds has increased. This is particularly important because the toxic VOCs may also be present at the workplace, in the classrooms, and at public places [7-9].

In our previous work [1], we demonstrated use of sensor arrays in detecting air pollution caused by common household products. The sensor arrays consisted of four to eleven sensors. The system was tested with seven household products, and it was able to detect the presence of tested products. Building on the outcome of the previous work, the presented study classifies the sources of the VOCs. In addition, this study was conducted using test-chambers and ultra-clean air. PCA was used as the pattern analysis technique.

The system development method, testing and analysis procedures, obtained results, and data analysis are presented and discussed.

II. SYSTEM DEVELOPMENT, ANALYSIS METHODS, AND TEST PROCEDURES

A. Sensor array

The developed array consists of eight sensors including seven cross-selective VOC sensors (Sensor 1 through Sensor 7) and a temperature sensor (Sensor 8). In order to choose appropriate sensors, some considerations of selection are required for these sensors, including sensitivity [4], response time, selectivity, and stability. While selecting the sensors to include in the array, we considered the fact that adding more sensors does not always provide additional useful information. More sensors also imply a higher cost for the array and the need for increased computational resources for analyzing the sensor data. Therefore, from initial selection of a larger number of sensors, those identified to contribute to the recognition were included in the sensor array.

B. Principal Component Analysis

PCA is a statistical dimensionality reduction technique where the new axes are selected for maximum covariance [11]. The covariance matrix can be calculated using equation (1).

$$\mathrm{cov}(X,Y)=\sum_{i=1}^{n}\frac{(X_i-X)(Y_i-Y)}{(n-1)} \tag{1}$$

Next, eigenvectors and eigenvalues of the covariance matrix are determined. The covariance matrix is a square matrix, denoted by A in equation (2). The non-zero vector x is called an eigenvector and the scalar λ is called an eigenvalue. The eigenvalues will be sorted by their magnitude shown in the equation (3). The magnitude of λ_1 is the largest and its corresponding eigenvector is called x_1, or the first principal component (PC1). Similarly, the second principal component (PC2) is found by finding the second eigenvector of the covariance matrix, x_2. The second eigenvector is perpendicular to the first PC and maximizes the covariance.

$$Ax=\lambda x \tag{2}$$

$$|\lambda_1|\geq|\lambda_2|\geq|\lambda_3|\geq\ldots\geq|\lambda_n| \tag{3}$$

The main goal of PCA is to convert a high dimensional multivariate data into a lower dimensional picture by using projections [12]. For example, the aforementioned PC1 and PC2 are the desired x and y-axes in a two-dimensional PCA plot. Thus, PCA enables simplification of a complex data set into a form that is easier to appreciate the relationship between different variables. In the presented study, PCA was implemented in MATLAB. The original data collected from the smart sensing system was imported into MATLAB and the presented results were generated.

C. Training and Testing for Identification

The method of identification for smart sensing systems involves two steps, known as training and testing. Training refers to a step of collecting and recording sensor data for the known compounds followed by analysis using pattern analysis techniques. While collecting data, it is important to keep the consistency for each trial so that the training data is not affected by external factors. Testing refers to the step of using the classification results from training to classify a test data set.

D. Experimental Setup and Data Collection

The schematic of the experimental setup is shown in Figure 1. It consists of three chambers, a control valve, and an ultra-clean air tank. The test VOCs were introduced in Chamber 1 with the air valve closed. Sensor array was placed in Chamber 2. Chamber 3 was used for trapping the outgoing VOCs. The air control valve was used to control the flow rate of the VOCs from Chamber 1 to Chamber 2. Glass wool was used in Chamber 1 to smoothen the VOC release process. The inner volume of all three chambers were 250 mL each. In order to develop an effective procedure, setup calibration testing was conducted with acetone and ethanol. These two chemicals were tested individually. Acetone or ethanol was injected in the first chamber, and various wait-times (time duration before opening the air-flow) and various flow-rates, were recorded to determine the parameters that resulted in the highest response from the sensor array. It was found that 120 second wait-time and 0.5 L/min were optimum parameters. Collection of the data began 45 seconds after opening the airflow valve and was collected for eight minutes. Purging was carried out for five minutes at 1.0 L/min. A temperature sensor was added in order to measure ambient temperature and ensure a stable temperature condition, and account for any variations. The sensor array was connected to an Arduino Mega to record and import data.

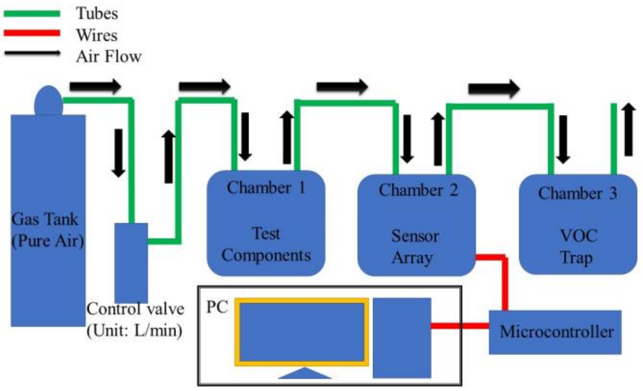

Fig. 1. Schematic diagram showing the test setup.

III. RESULTS AND DISCUSSION

In this study, we aim to classify various pollutants based on the collected sensor responses. From a list of common household products that are known to contribute to the indoor air pollution, the following representative compounds were selected for testing: Air Freshener (AF), Insect Repellent (IR), Lighter Fuel (LF), Ant Control (AC), Cleaner (CL), Paint (PA) and Paint Stripper (PS). Two sets of experiments consisting of tests with individual compounds (Section 3.1) and tests with a mixture of two or more compounds (Section III.B) were conducted.

A. Individual Compounds

A total of 30 sets of tests were conducted over an 11 day period. Sensor response data was collected for eight minutes for each test. A sample sensor array data for the Paint Stripper is shown in Figure 2. Sensors 1 to 8 corresponds to the eight sensors on the array.

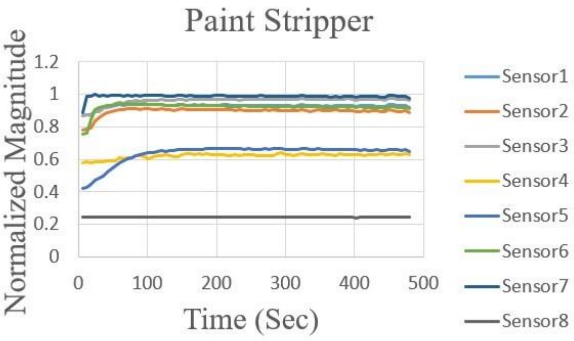

Fig. 2. A sample sensor array response raw data from a test with Paint Stripper.

For PCA analysis, three steady-state data points at 3, 4, and 5 minutes were sampled from each test. Thus, 24 data points were created from each test. PCA of all data sets are shown in Figure 3.

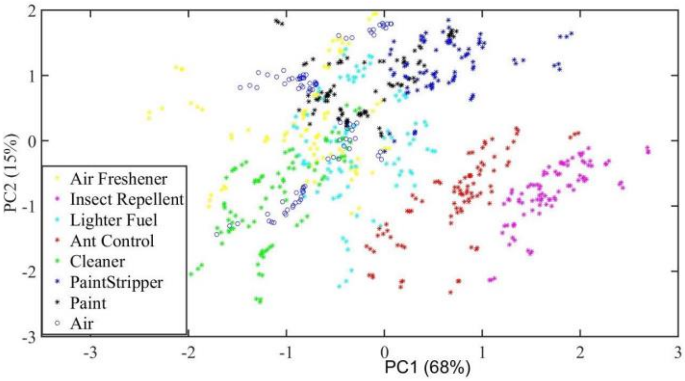

Fig. 3. PCA of training data set for seven household products, showing the first and the second principal components.

- Training and Testing

For testing purposes, each data set (containing data sampled at 3, 4, and 5 minutes) was tested against a training data set comprising of the remaining data sets (total of 87 data points). The process was repeated until each set was used as a test set. Table I shows the confusion matrix.

TABLE I. CONFUSION MATRIX

True Label	**Predicted Label (%)**							
	AF	***IR***	***LF***	***AC***	***CL***	***PS***	***PA***	***CA***
AF	99.74	0	0.04	0	1.15	0	0.04	0.73
IR	0	100	0	2.25	0	0.01	0	0
LF	0.01	0	99.21	0.01	0.02	0	0.01	37.45
AC	0	0	0.02	97.53	0	0.02	0	0.02
CL	0.16	0	0.01	0	98.72	0	0	0.09
PS	0	0	0.03	0.21	0	99.96	0	0.18
PA	0.02	0	0.22	0.01	0.01	0.01	99.92	5.01
CA	0.08	0	0.47	0	0.09	0	0.02	56.52

B. *Mixtures of Compounds*

Next, tests were conducted with a mixture of two or more compounds. The tested combinations are as follows: Insect Repellent and Ant Control; Paint and Paint Stripper; and Insect Repellent, Ant Control, Paint Stripper, and Paint. In addition, varied concentrations of the compound mixtures were tested. For example, a mixture of raised concertation of Insect Repellent and normal concentration of Ant Control , vice versa. Total of 30 sets of tests were conducted over a 4 day period. Sensor response data was collected for eight minutes for each test. The PCA results of all data sets are shown in Figure 4.

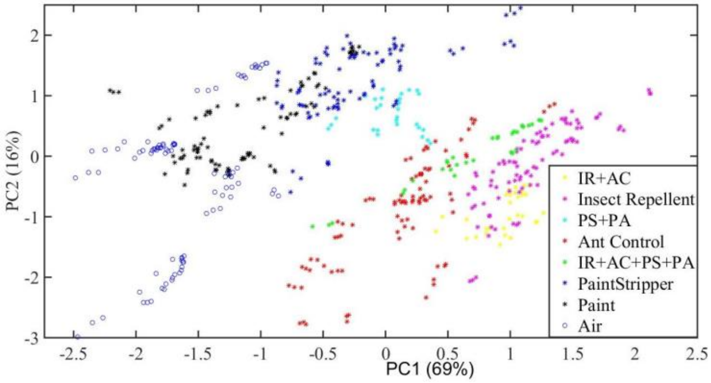

Fig. 4. PCA training data set for test with multiple VOC sources, showing first and second principal components.

The classification results of the data could be derived into a different pattern and presented as clusters in the PCA plot. However, the test results could be either Insect Repellent or Ant Control when these two items were mixed. This result matches a general prediction. While two chemicals were used, we could sense those two products. Likewise, while the mixture of paint and paint stripper were tested, the recognition results would be either Paint or Paint Stripper. For testing, procedure of using each data set as a test against a training data set comprising of the remaining data sets was utilized. The confusion matrix of the results is shown in Table II.

TABLE II. THE LIKELIHOODS IDENTIFYING RESULTS

True Label	**Predicted Label (%)**				
	IR+AC	***PS+PA***	***IR+AC+ PS+PA***	***IR(higher)+AC(normal)***	***AC(higher)+IR(normal)***
AF	0	0	0.01	0.04	0
IR	95.61	0.05	54.57	28.63	94.24
LF	0	0.02	0.06	0.15	0
AC	4.34	1.85	42.09	69.81	5.64
CL	0	0	0	0.02	0
PS	0.05	98.01	3.14	1.16	0.12
PA	0	0.07	0.1	0.16	0
CA	0	0	0.01	0.04	0

In order to improve the prediction from above results and to produce an actionable output for the users, likelihood results as shown in Table III were developed. The likelihood was calculated based on the Euclidean distance between the two groups. Euclidean distance is used to describe the distance between two dots in Euclidean space. The Euclidean distance formula is shown in equation (4) where “d” represents the distance. The parameters p_1 to p_8 are the mean response values of the test data set and q_1 to q_8 are from the training data set, respectively.

$$d(p,q)=\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\ldots+(p_n-q_n)^2} \tag{4}$$

The similarity is higher while the distance between two groups are shorter. Thus, the likelihood can be derived into equation (5). The summation of (1/d) represents the reciprocal of the summation distances between test data and the eight training products. However, in this case, we decided to use eight sensors such that each piece of data is treated as an eight-dimensional dot. If we simply implement equation (5) to generate the probability, the percentages would likely be too small. Thus, we have to take into account the number of sensors and normalize the equation (5). Equation (5) can be rewritten as equation (6), which gives the desired probability.

$$p = \frac{1/d}{sum(1/d)} \tag{5}$$

$$p_{nor} = \frac{1/d^8}{sum(1/d^8)} \tag{6}$$

Another set of experiments tested two of the products were mixed with different concentration levels. First, 0.6 mL IR mixed with 0.2 mL AC and second, 0.2 mL IR mixed with 0.6 mL AC (Table II). Based on the classification results, enhancing the concentration only change the intensity, and it is was not very effective for identification. Creating another mixture recognized as a totally different cluster of data because of the differences for sensitivities toward different components. When the concentration of Insect Repellent was enhanced, the results were more likely to be an Ant Control. While the concentration of Ant Control was enhanced, the likelihood of Insect Repellent was higher. In this case, the reasonable understanding would be the new mixture was slightly similar or near to their original pattern of the components which caused the high accuracy. Changing the concentrations causes a stronger response and affects the response values of the sensing system, but does not affect the test results.

TABLE III. SIMILARITY INDEX TABLE

	Similarity	
	Similarity Description	***Percentage (%)***
1	Very unlikely	0-25
2	Unlikely	25-50
3	Not very likely	50-70
4	Likely	70-90
5	Very likely	Above 90

IV. CONCLUSIONS

We have presented and discussed a system for detecting indoor air VOCs. The system consisted of a chemical sensor array and PCA classifier. The system was able to classify common household sources of harmful VOCs. However, additional training data is needed for the system to be able to classify a wide range of VOCs and their sources. We believe that the presented system makes significant contributions towards developing a modality for passive and continuous monitoring of indoor air VOCs using chemical sensor array. Compared to our previously reported work, which used filtered air and open test chambers, in the presented study, air-tight chambers set up with ultra-clean air was utilized. Similarly, more controlled calibration methods were utilized to determine the air-flow and test procedures. In addition, much extensive tests and results analysis are presented in this study, including, tests and analysis with multiple VOCs. The information about possible sources of the VOCs will help the users to take corrective measures.

ACKNOWLEDGMENT

The authors would like to thank Dr. Indrs Sahu in the Department of Chemistry at Miami University for his laboratory assistance. The authors would also like to thank Dr. Chi-Hao Cheng, Mr. Matt Boubin, Mr. Tanner McWhorter, Mr. Jeff Peterson, Ms. Tina Carico, and Dr. Qihou Zhou in the Department of Electrical and Computer Engineering for their assistance with this work.

REFERENCES

[1] Y.-T. Chen, Z. Samborsky, and S. Shrestha “Electronic nose for ambient detection and monitoring,” in *Proceedings of SPIE Commerical Scientific and Imaging*, 2017, pp. 1-8.

[2] “Volatile Organic Compounds' Impact on Indoor Air Quality,” *EPA*, 06-Nov-2017. [Online]. Available: https://www.epa.gov/indoor-air-quality-iaq/volatile-organic-compounds-impact-indoor-air-quality.

[3] “Health Effects due to Indoor Air Pollution,” *Indoor and Built Environment*. [Online]. Available: http://journals.sagepub.com/doi/abs/10.1177/1420326X03037109.

[4] A. P. Jones, *"Indoor air quality and health," Atmospheric Environment, vol. 33, no. 28, pp. 4535–4564, 1999. - Open Access Library*.[Online]. Available: http://www.oalib.com/references/14525399.

[5] V. Mishra and R. Agarwal, “Sensitivity, response and recovery time of SnO2 based thick-film sensor array for H2, CO, CH4 and LPG,” *Microelectronics Journal*, vol. 29, no. 11, pp. 861–874, 1998.

[6] P. Wolkoff and G. D. Nielsen, “Organic compounds in indoor air—their relevance for perceived indoor air quality?,” *Atmospheric Environment*, vol. 35, no. 26, pp. 4407–4417, 2001.

[7] J. M. Daisey, W. J. Angell, and M. G. Apte, “Indoor air quality, ventilation and health symptoms in schools: an analysis of existing information,” *Indoor Air*, vol. 13, no. 1, pp. 53–64, 2003.

[8] S. Lee and M. Chang, “Indoor and outdoor air quality investigation at schools in Hong Kong,” *Chemosphere*, vol. 41, no. 1-2, pp. 109–113, 2000.

[9] E. Llobet, J. Brezmes, X. Vilanova, J. E. Sueiras, and X. Correig, “Qualitative and quantitative analysis of volatile organic compounds using transient and steady-state responses of a thick-film tin oxide gas sensor array,” *Sensors and Actuators B: Chemical*, vol. 41, no. 1-3, pp. 13–21, 1997.

[10] “Indoor Air Quality (IAQ),” *EPA*, 14-Mar-2018. [Online]. Available: https://www.epa.gov/indoor-air-quality-iaq.

[11] J. Shlens ,“A Tutorial on Principal Component Analysis” ,07-Apr-2014. [Online]. Available: https://arxiv.org/pdf/1404.1100.pdf

[12] C. D. Natale, F. Davide, and A. Damico, “Pattern recognition in gas sensing: well-stated techniques and advances,” *Sensors and Actuators B: Chemical*, vol. 23, no. 2-3, pp. 111–118, 1995.

[13] S. Saad, A. Andrew, A. Shakaff, A. Saad, A. Kamarudin, and A. Zakaria, “Classifying Sources Influencing Indoor Air Quality (IAQ) Using Artificial Neural Network (ANN),” *Sensors*, vol. 15, no. 5, pp. 11665–11684, 2015