https://doi.org/10.2352/ISSN.2470-1173.2020.15.COLOR-353
© 2020, Society for Imaging Science and Technology

Using Acoustic Information to Diagnose the Health of A Printer*

Chin-Ning Chen[a], Katy Ferguson[b], Anton Wiranata[b], Mark Shaw[b], Wan-Eih Huang[a], George Chiu[a], Patricia Davies[a], and Jan P. Allebach[a]
[a]School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47906, U.S.A.
[b]HP Inc., Boise, ID 83714, U.S.A.

Abstract

We present a sound-based anomaly detection system to diagnose printer health. Also, we improve the model performance by using acoustic data augmentation. We first use the detector to extract the important acoustic information from the input printer sound. Second, we use principal component analysis to do feature extraction. Third, we feed the extracted features from the previous step into the two different anomaly detection models to evaluate the model performances. Finally, we go through the same system pipeline with different augmented training data to see whether or not acoustic data augmentation can improve the model performance.

1. Introduction

Anomaly detection is used in a variety of applications, such as fraud detection for credit cards, security systems, and machine operational conditions [1]. Also, the application of sound-based systems have been attracting more attention because of inexpensive microphone settings and recordings [24] . In this paper, we combine these two concepts to model a sound-based anomaly detection system for a printer.

Recent researches have already explored plenty of techniques for modeling the acoustic signal in order to better capture its important features. Various hand-crafted descriptors have been proposed such as mel frequency cepstral coefficient (MFCC) [2], filter bank [3, 4], spectrogram [5, 16], and bag-of-audio-words [7,8]; and they were modeled with Support Vector Machine (SVM) [9]. But even though we have multiple feature representations to fit the specific application such as automatic speech recognition (ASR) [10] and acoustic scene classification (ASC) [11], there still has one problem: the lack of the acoustic data. Unlike image datasets, there are few public datasets that are suitable for various acoustic applications. As a result, based on the limited data, data augmentation plays a critical role to expand the dataset size.

A variety of acoustic data augmentation methods have been proposed as follows. Vocal tract length normalization (VTLN) [12] transforms the spectrogram using a random linear warping along the frequency dimensions. [13] uses the modified version of [12] with a fixed gap of the warping factor. [14] proposes Equalized Mixture Data Augmentation (EMDA) to augment the sound by randomly mixing two sounds of a class, with randomly selected timings. Furthermore, this method perturbs the sound by amplifying/attenuating a particular frequency band. Similarly, [15] makes the assumption that a combination of two or more audio segments from the same scene is another sample of that scene with more complex pattern and events. Also, [16] mixes training samples together to do augmentation. Instead of mixing audio sequences, [17] inserts blank rectangles into the two-dimensional Mel-spectrogram with a randomly chosen size and location to remove some information as their augmentation method.

*Research supported by HP Inc., Boise, ID 83714

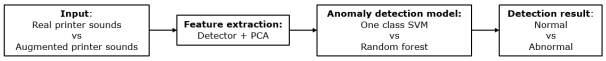


Figure 1: System pipeline for printer sound-based anomaly detection.

Figure 1 shows the system pipeline of our work. Our goal is to use an acoustic signal to diagnose the printer health. Also, we would like to see whether or not the augmented printer sounds can improve the anomaly detection. In this paper, first of all, the extracted features are based on the detector [18] and principal component analysis (PCA) [19]. Second, the anomaly detection models that we use are one class support vector machine (OCSVM) [20] and random forest (RF) [21]. Third, we use three augmentation methods: pitch shifting, time stretching from [22], and the mix-up concept. Notice that we will train the classifiers with augmented datasets and test them with the real collected printer sound.

2. Proposed method

The proposed method consists of four parts. The first two parts are feature analysis of the first stage feature extraction with the detector and the design of defect generator. The third part is the second stage feature extraction with PCA. The last part is the introduction of the anomaly detection model that we are using.

2.1 Feature Analysis

In feature analysis, first of all, we will introduce the detector, which is used for the first stage feature extraction. Second, based on the feature distribution of the normal real printer sounds, we can define what kind of features represent an abnormal characteristic. Third, based on the feature distribution that we just defined in the second part, we can artificially synthesize the abnormal printer sounds.

2.1.1 Detector

We use the detector based on [18] as our first stage feature extraction. Figure 2 shows the pipeline of the detector. Basically, it is constructed in two parts by strong tone information and modulation information.

For the strong tone information, we first calculate the power spectrum density (PSD) of the input printer sound. Next, we use a moving average filter to find the dynamic threshold to extract the strong tone frequencies, relative PSD, absolute PSD, and peak

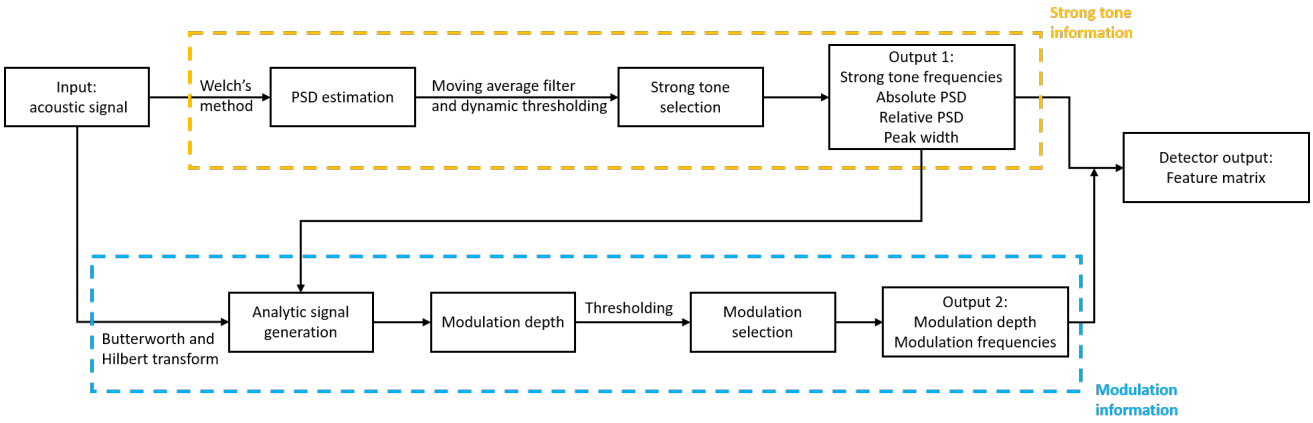


Figure 2: Detector pipeline.

width.With this information for each strong tone, we can step to the modulation information part to generate the analytic signal by using a Butterworth filter and the Hilbert transform. Based on the generated analytic signal, we can calculate its modulation depth and find its corresponding modulation frequencies. The final output of the detector, which includes strong tone and modulation information, is called the feature matrix as shown in Figure 3.

Strong Tone Frequency (Hz)	Relative PSD	Absolute PSD	Peak Width	Modulation Frequency (Hz)	Modulation Depth (%)
1992	26.72548	33.69137	4	0	0
11895	23.43187	24.94561	3	1	10.13539
13148	21.59192	22.21689	3	0	0
14414	15.23058	17.20906	4	1	45.88093
14414	15.23058	17.20906	4	2	12.84543
352	14.68129	23.71069	3	0	0

Figure 3: Example of the feature matrix.

2.1.2 Feature Distribution

Because we only have collected real normal printer sounds, we would like to find some characteristic of abnormal features based on the analysis of normal data. Figure 4 shows the histogram of the strong tone frequencies from the collected normal printer sounds. Based on the characteristic shown in Figure 4, we will synthesize the synthetic abnormal printer sounds. We first use two normal distributions to fit the strong tone frequency histogram as shown in Figure 5 and then define the frequency ranges that represent abnormal features. Based on the observation of Figure 3 and 5, the definition of the abnormal features are: strong tone frequency ranges from 3 kHz to 10 kHz and modulation depth larger than 100%.

2.1.3 Defect Generator

The purpose of the defect generator is to generate the synthetic abnormal acoustic signal since we don't have real abnormal printer sounds. Based on the abnormal features that we defined and the concept of amplitude modulation (AM), we can specify certain abnormal strong tone frequencies as the carrier signal to carry the modulation frequency as the modulating signal.

Take Figure 6 as an example, we specify the strong tone frequency at 5 kHz as the carrier signal to carry the modulating signal in Equation 1 and 4, respectively, where $m_1(t)$ is the simulated square wave based on a Fourier Series representation, as shown in

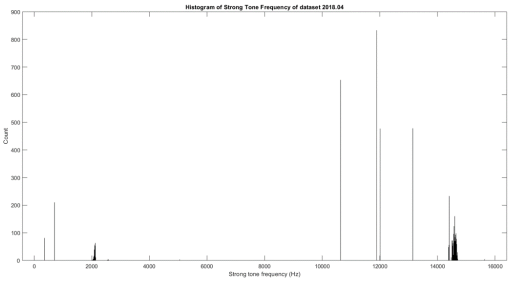


Figure 4: Histogram of strong tone frequencies from real normal printer sound dataset.

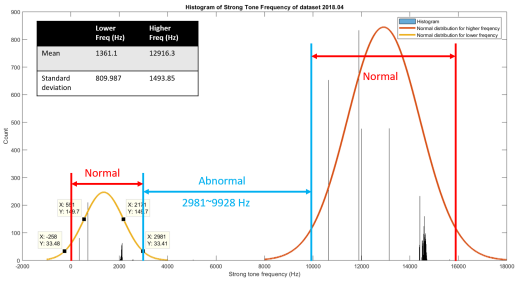


Figure 5: Fitting histogram of strong tone frequencies with two normal distributions.

Equation 2 and $m_2(t)$ is the carried modulation frequency at 5 Hz in Equation 3.

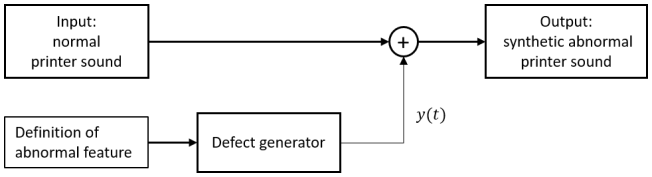
$$c(t) = A_c \cdot sin(2\pi 5000t) \tag{1}$$

$$m_1(t) = \frac{1}{2} + \sum_{k=1}^{19} \frac{2}{\pi k} sin(2\pi kt) \tag{2}$$

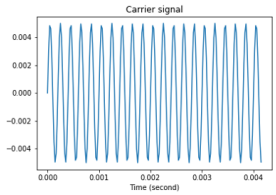
$$m_2(t) = \frac{2}{\pi} \cdot sin(2\pi 5t) \tag{3}$$

$$m(t) = m_1(t) + m_2(t) \tag{4}$$

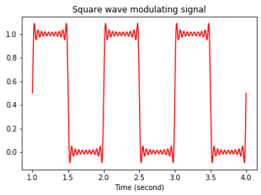
$$y(t) = (1 + m(t))c(t) \tag{5}$$



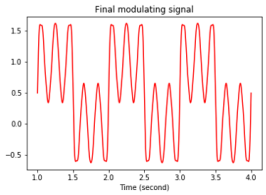
(a) Defect generator pipeline



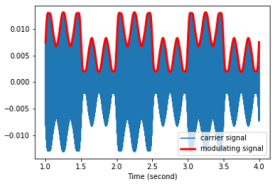
(b) Carrier signal $c(t)$



(c) Simulated square wave $m_1(t)$



(d) Final modulating signal $m(t)$



(e) Final defect $y(t)$

Figure 6: Example of defect generator with strong tone frequency at 5 kHz carrying modulation frequency at 5 Hz.

2.2 Feature Extraction

We have already shown that the first stage feature extraction is the output of the detector called the feature matrix. The column dimension of the feature matrix is fixed at 6. But the row dimension of the feature matrix varies for different printer sounds according to the detector algorithm output. As a result, based on this feature matrix, we can further extract a fixed-length feature vector during the intermediate steps in the PCA to represent each printer sound. To do this, each printer sound has to go through the following steps:

Step 1: Based on the extracted feature matrix as shown in Figure 3, we normalize each column into unit length with Euclidean norm.

Step 2: Based on the normalized feature matrix, we can find its mean vector $\vec{\mu}$ and covariance matrix Σ as shown by Equations 6 and 7, respectively. In our work, $\vec{x}_i$ is the i-th row feature vector within the feature matrix.

$$\vec{\mu} = \frac{1}{n}\sum_{i=1}^{n}\vec{x}_i \tag{6}$$

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(\vec{x}_i - \vec{\mu})^T(\vec{x}_i - \vec{\mu}) \tag{7}$$

Step 3: Find all of the eigenpairs, eigenvalues, and their corresponding eigenvectors, of the covariance matrix that satisfy the Eigen-equation 8.

$$\Sigma\vec{w} = \lambda\vec{w} \tag{8}$$

Step 4: Sort the eigenvalues in descending order and their corresponding eigenvectors.

Step 5: Finally, we choose the eigenvector that corresponds to the largest eigenvalue, which is also called the first principal component, as our final feature to feed into our anomaly detection model.

All of the audio files have to go through the same process from Step 1 to Step 5. Even though the detector extracts feature matrices with different dimensions from different audio files, we still can find a feature with fixed dimension to represent each audio file.

There are two differences between the normal PCA process and our PCA feature. First of all, people normally use PCA on all the acoustic signals at one time. But in our case, we use PCA separately on the feature matrix for each input acoustic signal. Second, people normally use the reduced feature to do further processing. But in our case, we use the first principal component as our feature to do further processing. However, in both cases, we can find the features with fixed dimension to represent each acoustic signal.

2.3 Anomaly Detection Model

We use the semi-supervised classifier OCSVM and the supervised classifier RF as our anomaly detection models.

2.3.1 One Class Support Vector Machine

The concept of OCSVM [20] is that it maps input data into a high dimensional feature space via a kernel and finds the maximal margin hyperplane which best separates the training data from the origin in the mapped feature space. In mathematical terms, OCSVM is solving the optimization problem stated in Equation 9 [20]

$$\begin{aligned}&\min_{\mathbf{w},\xi_i,\rho}\ \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{vl}\sum_{i=1}^{l}\xi_i - \rho \\ &\text{subject to } \left(\mathbf{w}^T\Phi(\mathbf{x}_i)\right) \le \rho - \xi_i,\ i = 1,\ldots,l,\ \xi_i \ge 0\end{aligned} \tag{9}$$

where

$$\begin{aligned}&\mathbf{w} = \sum_i \alpha_i \cdot \Phi(\mathbf{x}_i) \\ &\sum_i \alpha_i = 1\end{aligned} \tag{10}$$

The decision function is

$$f(\mathbf{x}) = sign\left(\mathbf{w}^T\Phi(\mathbf{x}_i) - \rho\right) \tag{11}$$

where $v \in (0,1]$ is the fraction of outliers, l is the number of data points, ξ_i is the slack variable, ρ is an offset parameter associated with the kernel, and $\Phi(\cdot)$ is the kernel function that maps the training sample into another space. If the data are linearly separable, then a linear kernel will work well. However, if the data is not linearly separable, then a non-linear kernel should be used. Here, we use a non-linear kernel, namely the radial basis function. The way we categorize the data as normal or abnormal is based on the decision function. If $f(\mathbf{x}) > 0$, we label $\mathbf{x}$ as normal, and if $f(\mathbf{x}) < 0$, we label $\mathbf{x}$ as abnormal.

2.3.2 Random Forest

Random forest [21] is a supervised ensemble classifier, which is constructed by a set of multiple decision trees. The more trees it has, the more robust the forest is. Each decision tree will report class prediction. Finally, we follow the wisdom of crowds rule to classify the input to the class with the most votes. We also use the bagging (bootstrap aggregation) method [21] to randomly extract a subset of the features in each decision tree.

3. Experimental Evaluation

In this section, we first introduce the dataset and the data arrangement for the anomaly detection model. Next, we show the parameter settings for the data augmentation. Finally, we show the evaluation results based on different augmented datasets and different classifiers.

3.1 Datasets

We use our collected 400 real normal printer sounds dataset as a comparison basedline. We additionally synthesize 100 abnormal printer sounds. Among these 100 synthetic abnormal printer sounds, all of them are used for RF and 40 of them are used for OCSVM. These 400 collected normal plus 40 synthesized abnormal printer sounds are the data arrangement for OCSVM and 400 collected normal plus 100 synthesized abnormal printer sounds are the data arrangement for RF.

For OCSVM, we randomly choose 360 within the 400 real normal printer sounds as the normal training data. The remaining 40 real normal printer sounds and the 40 synthetic abnormal printer sounds are the testing data. For RF, we randomly choose 90 within the 100 synthetic abnormal printer sounds as the abnormal training data and randomly choose 360 within the 400 real normal printer sounds as the normal training data. The remaining 40 real normal and 10 synthetic abnormal printer sounds are the testing data. Note that for RF, we have fewer synthetic abnormal sounds for testing, because we need the rest of them for training. Here, all of the evaluation results are based on 10-fold cross validation.

3.2 Data Augmentation

We use the previously mentioned three augmentation methods to generate four augmented datasets: larger/smaller pitch shifting datasets, time stretching dataset, and mixture dataset. For larger and smaller pitch shifting, we pitch shift with values $\pm 1, \pm 2$ and $\pm 0.1, \pm 0.2$, respectively. And for time stretching, we stretch with values $\pm 0.01, \pm 0.02$. For mixture, first, we randomly choose two audio files from the real printer sound dataset. Next, in the time domain, we combine them as shown in Equation 12.

$$x_n = 0.5 \cdot x_i + 0.5 \cdot x_j \tag{12}$$

where x_n is the augmented audio file and x_i and x_j are the two randomly chosen audio files from the same class. The augmented result keeps the same label as its mixture source.

3.3 Results

Figure 7 and 8 show the accuracy as a function of the number of the normal training data. For both OCSVM and RF, the number of normal training data ranges from 80 to 360; and at each data point, we increase the size of the set by 40 normal training data. For OCSVM, because it is semi-superviesd learning classifier, we only need normal training data. As a result, the number of synthetic abnormal training data for OCSVM is zero. For RF, the number of synthetic abnormal data ranges from 20 to 90; and at each data point, we increase the size of the set by 10 abnormal training data. Table 1 shows the results of accuracies achieved with different combinations of augmentation method and classifier when we have the largest set of normal training data, which is 360, and 90 abnormal training data. The value inside the parentheses is the standard deviation. Based on Figures 7 and 8, and Table 1, we can see that both pitch shifting and time stretching perform better than the real printer sound dataset with OCSVM. Also, the augmented dataset shows obvious improvement with OCSVM, but performs similarly to the real printer sound dataset with RF. Note that the evaluation is based on the same set of testing data for all cases.

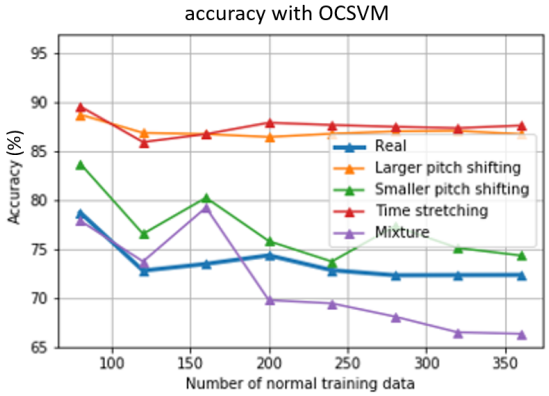


Figure 7: Accuracy with OCSVM as a function of the number of the normal training data.

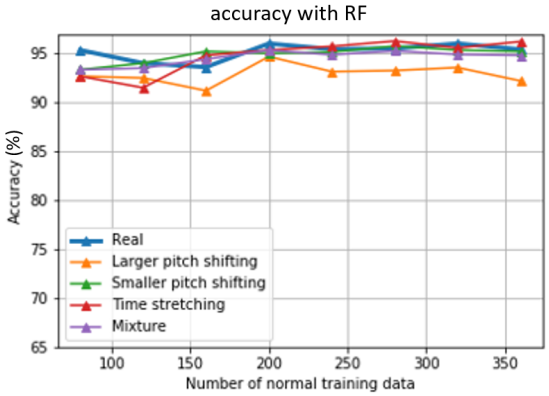


Figure 8: Accuracy with RF as a function of the number of the normal training data.

4. Conclusion

We have developed an anomaly detection system to diagnose printer health. Our proposed anomaly detection pipeline consists of the following steps: First, data preparation for synthetic abnormal sounds and the augmented sounds. Second, feature extraction based on a detector and principal component analysis. Third, categorize the input printer sound into the normal or abnormal class. Our results show that the augmented dataset can

	Classifiers	
Training dataset	OCSVM (%)	RF (%)
Real	72.38 (1.89)	95.4 (2.54)
Larger pitch shifting	86.75 (2.03)	92.2 (2.89)
Smaller pitch shifting	74.38 (2.45)	95.2 (2.71)
Time stretching	**87.63** (3.33)	**96.2** (1.66)
Mixture	66.38 (2.13)	94.8 (2.56)

Table 1: Comparison classification accuracies achieved with different combinations of augmentation method and classifier.

improve the performance of our anomaly detection model, especially for OCSVM. We are continuing to investigate novel augmentation methods and different classifiers to improve the model performance, and find the most appropriate one.

References

[1] Y. Ono, Y. Onishi, T. Koshinaka, S. Takata, and O. Hoshuyama, "Anomaly detection of motors with feature emphasis using only normal sounds," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2800-2804, 2013

[2] H. Phan, L. Hertel, M. Maass, R. Mazur, and A. Mertins, "Representing nonspeech audio signals through speech classification models," in *International Speech Communication Association (ISCA)*, pp. 3441-3445, 2015

[3] W. Choi, S. Park, D. K. Han, and H. Ko, "Acoustic event recognition using dominant spectral basis vectors," in *International Speech Communication Association (ISCA*), pp. 2002–2006, 2015

[4] J. Beltrán, E. Chávez, and J. Favela, "Scalable identification of mixed environmental sounds, recorded from heterogeneous sources," in *Pattern Recognition Letters*, vol. 68, pp. 153-160, 2015

[5] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8614-8618, 2013

[6] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with time–frequency audio features," in *Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 1142–1158, 2009

[7] S. Pancoast and M. Akbacak, "Bag-of-audio-words approach for multimedia event classification," in *International Speech Communication Association (ISCA)*, pp. 2105–2108, 2012

[8] H. Lim, M. J. Kim, and H. Kim, "Robust sound event classification using LBP-HOG based bag-of-audio-words feature representation," in *International Speech Communication Association (ISCA)*, pp. 3325–3329, 2015

[9] X. Lu, P. Shen, Y. Tsao, C. Hori, and H. Kawai, "Sparse representation with temporal max-smoothing for acoustic event detection," in *International Speech Communication Association (ISCA)*, pp. 1176–1180, 2015

[10] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4277–4280, 2012

[11] Y. Han, J. Park, and K. Lee, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," in *Detection and Classification of Acoustic Scenes and Events (DCASE)*, pp. 1–5, 2017

[12] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, 2013

[13] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," in *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, pp. 1469–1477, 2015

[14] N. Takahashi, M. Gygli, B. Pfister, and L. V. Gool, "Deep convolutional neural networks and data augmentation for acoustic event detection," in *arXiv preprint arXiv:1604.07160*, 2016

[15] H. Zeinali, L. Burget, and J. Cernocky, "Convolutional neural networks and x-vector embedding for DCASE2018 acoustic scene classification challenge," in *arXiv preprint arXiv:1810.04273*, 2018

[16] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *arXiv preprint arXiv:1710.09412*, 2017

[17] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *arXiv preprint arXiv:1708.04896*, 2017

[18] X. Xue, N. Kim, X. Wang, J. Allebach, J. S. Bolton, G. Chiu, P. Davies, K. Ferguson, D. Kaisle, and M. Shaw, "Digital signal processing for laser printer noise source detection and identification," in *The Proceedings of Noise-Con*, 2019

[19] S. Raychaudhuri, J. M. Stuart, and R. B. Altman, "Principal components analysis to summarize microarray experiments: application to sporulation time series," in *Biocomputing*, pp. 455–466, 1999

[20] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," in *Neural Computation*, vol. 13, pp. 1443–1471, 2001

[21] L. Breiman, "Random forests," in *Machine learning*, vol. 45, pp. 5–32, 2001

[22] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," in *IEEE Signal Processing Letters*, vol. 24, pp. 279–283, 2017

[23] S. Y. Kung, "Kernel methods and machine learning," in *Cambridge University Press*, 2014

[24] T. Hayashi, T. Komatsu, R. Kondo, T. Toda, and K. Takeda, "Anomalous sound event detection based on WaveNet," in *IEEE European Signal Processing Conference (EUSIPCO)*, pp. 2494-2498, 2018

Author Biography

Chin-Ning Chen received both her BS and MS in electrical and computer engineering from the National Chiao Tung University, Hsinchu, Taiwan, in 2014 and 2016, respectively. Currently, she is a Ph.D. student in Purdue University, West Lafayette, IN, where she has been since 2016. Her research interests include machine learning, deep learning, and image/audio processing.

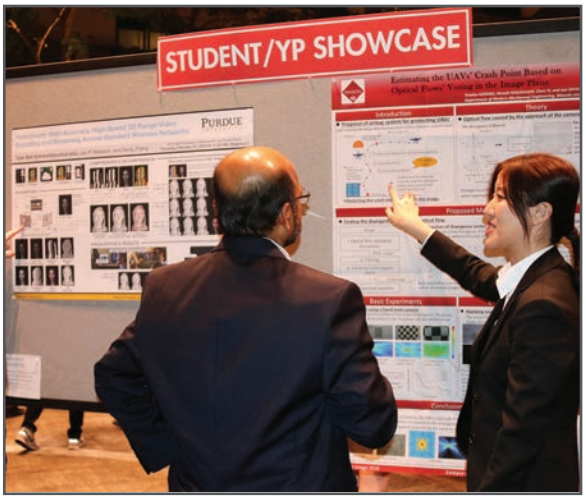
JOIN US AT THE NEXT EI!

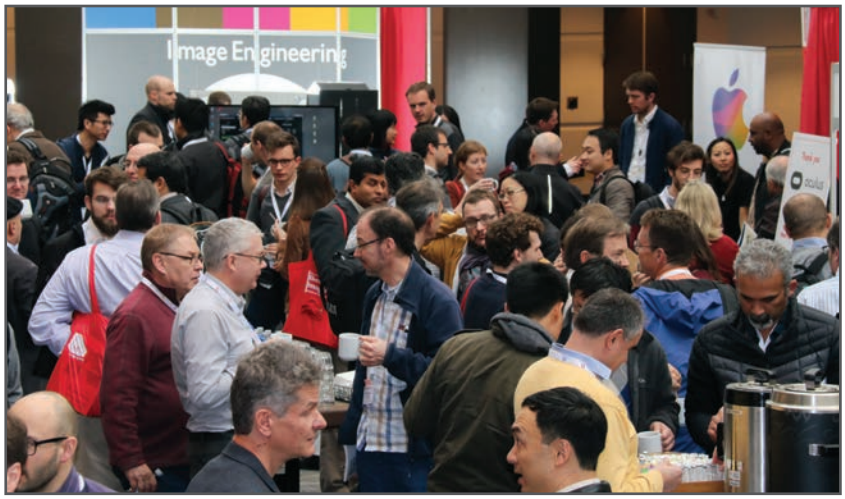
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org