Spectro-Temporal Analysis using Local Binary Pattern Variants for Acoustic Scene Classification

Shamsiah Abidin, *Student Member, IEEE,* Roberto Togneri, *Senior Member, IEEE,*
and Ferdous Sohel, *Senior Member, IEEE*

***Abstract*—In this paper we present an approach for acoustic scene classification, which aggregates spectral and temporal features. We do this by proposing the first use of the variable-Q transform (VQT) to generate the time-frequency representation for acoustic scene classification. The VQT provides finer control over the resolution compared to the constant-Q transform (CQT) or STFT and can be tuned to better capture acoustic scene information. We then adopt a variant of the local binary pattern (LBP), the Adjacent Evaluation Completed LBP (AECLBP), which is better suited to extracting features from acoustic time-frequency images. Our results yield a 5.2% improvement on the DCASE 2016 dataset compared to the application of standard CQT with LBP. Fusing our proposed AECLBP with HOG features we achieve a classification accuracy of 85.5% which outperforms one of the top performing systems.**

***Index Terms*—acoustic scene, local binary patterns, feature extraction, time-frequency analysis, fusion**

I. Introduction

THE research on acoustic scene classification has been of interest to researchers in the area of acoustic analysis for the past two decades. Acoustic scene analysis has been used in applications such as automatic audio surveillance, mobile phone sensing, context-aware assistive robots, music genre classification and multimedia archiving. Audio surveillance is one of the applications that typically employs sound content analysis techniques to detect outlier activities such as gunshot and screaming in a specific indoor environment [1], [2]. Furthermore, in order to achieve the environment awareness, a mobile phone application is expected to be able to identify and automatically adapt to the surrounding environments [3], [4]. The objective of acoustic scene classification (ASC) is to identify the environment in which an audio stream has been produced [5]. The DCASE 2013 Challenge was introduced by the IEEE AASP Technical Committee to provide an evaluation and comparison of different techniques developed in acoustic scene classification on a benchmark dataset. Intended for inspiring the development of novel methodologies and improving the state-of-the-art in ASC, the DCASE 2016 Challenge dataset for audio scene classification was released with more challenging data.

Time-frequency representations (TFR) of discrete-time signals play an important role in acoustic analysis. A TFR provides a visual representation of the temporal and spectral structures that can be viewed as a 2D texture image. The constant-Q transform (CQT), commonly used for music processing tasks [6], [7], has now been applied to acoustic scene analysis [8], [9]. The use of CQT in combination with feature learning approaches based on nonnegative matrix factorization (NMF) by [9] has achieved a classification accuracy of 83.8% which is the state-of-the-art non-neural network based system on the DCASE 2016 dataset. However, the CQT lacks flexibility as the Q-factor is constant throughout the frequency band analysis. Indeed, a variable-Q factor is necessary to retain the important spectral and temporal structure of the acoustic signal. A Q-factor which is adaptable to the acoustic signal is required to produce a more accurate TFR representation.

The TFR texture image features can be extracted by well-known feature extraction methods used in computer vision. The Local Binary Pattern (LBP) is a state-of-the-art feature extraction method for analyzing image textures due to its computational simplicity [10]. A number of different variants of LBP have been developed to improve its robustness, and to increase its discriminative capability and applicability to different types of problems in image-classification applications. The Adjacent Evaluation LBP (AELBP) [11] is introduced to improve noise robustness of LBP by introducing the adjacent evaluation window and modifying the threshold scheme of LBP. It can be used with existing LBP variants such as the Completed Local Binary Pattern (CLBP) [12], the Completed Local Binary Count (CLBC) [13] and the Local Ternary Pattern (LTP) [14] to derive new image features against noise for texture classification. The CLBP feature extraction decomposes the image's local structure into two complementary components: the difference signs and the difference magnitudes to provide more discriminative information.

Motivated by the advancement of LBP variants, we are inspired to identify a variant of LBP that is adaptable to the acoustic signal representation and better suited for ASC. The micro structure of the image texture is different from the TFR. The image texture might have rotation and illumination variations whereas in the case of TFR there is no illumination or rotation variations. Also, in the case of TFR,

S. Abidin and R. Togneri are with the School of Electrical, Electronics and Computer Engineering, The University of Western Australia, Perth, WA, 6009 Australia e-mail: (shamsiah.abidin@research.uwa.edu.au; roberto.togneri@uwa.edu.au).

F. Sohel is with School of Engineering and Information Technology, Murdoch University, Perth 6150, WA, Australia e-mail: (f.sohel@murdoch.edu.au).

Manuscript received month date year; revised month date year

the random intensity of 'pixel' values might be considered as noise in LBP. The smoothing effect by using adjacent evaluation window should reduce the interference from the saltation of neighbors' values. Moreover, incorporating the magnitude component of the LBP in the TFR offers local contrast and variance information to extract more discriminative patterns, which is necessary for ASC.

In this paper, we propose a novel acoustic scene classification framework, which uses the unique combination of variable-Q transform (VQT) and Adjacent Evaluation Completed LBP (AECLBP) to provide state-of-the-art performance. A VQT aggregates local spectro-temporal features suitable for ASC as it has a smoothly varying Q-factor that captures the time resolution at lower frequencies better than the CQT [7]. The VQT representations were used for acoustic event detection (AED) by [15]. However in [15], the VQT used for generating the spectrogram and used by the proposed NMF, there was no mention of its superiority to the CQT. We present a new contribution by adapting AECLBP [11], to extract relevant features based on the VQT TFR for ASC. In AECLBP, an adjacent evaluation window around the neighboring pixel is constructed to modify the threshold scheme of LBP to address the issue of noise sensitivity in ASC. Also, it includes the information contained in the magnitudes of local differences as complementary to the signs of the traditional LBP. Since the acoustic scenes have large variations in time, each TFR is segmented along the frequency axis to extract local information, which better highlight distinctive spectral regions of the scene. We apply the zoning techniques as proposed in our previous work [16], in which the TFR is segmented into non-overlapping uniform slices. The AECLBP threshold values are computed distinctively for each zone to better capture the local intensity information. AECLBP histograms are computed for each zone and then concatenated. Finally, we present the results from a simple feature level fusion of Histogram of Oriented Gradients (HOG) and AECLBP features. The HOG features provide complementary information by providing the distribution of gradients at different orientations of the time-frequency images. In this work we use the standard SVM as the classifier. The experimental results show that the proposed framework outperforms the state-of-the-art systems on the DCASE 2016 dataset.

The remainder of the paper is organized as follows: In Section II, a review of relevant work is given. The proposed framework is explained in Section III, Section IV describes the experimental setup, results and analysis, and we provide our conclusions in Section V.

II. Previous Work

The analysis of environmental audio was pioneered by Sawhney and Maes in [17] who used the power spectral density and Gamma-tone filter-bank to mimic the response of the human cochlea. The studies of ASC initially focussed on spectral features, which have been adapted from speech recognition research such as zero crossing rate, frequency-band energy, spectral centroid, linear predictive coefficients and Mel-frequency cepstral coefficients (MFCC) [18], [19]. The MFCC features provide a short-time spectral analysis to generate a set of cepstral coefficients per analysis frame which is widely used for feature extraction. The combination of MFCC with Gaussian Mixture Models (GMM) has been a common approach for the baseline system for ASC [20], [21]. In the DCASE 2013 challenge, Roma et al. [22] outperformed all of the competing methods by employing recurrence quantification analysis (RQA) parameters to encode the series of MFCC coefficients, which were then classified by an SVM.

In current approaches to accurately discriminate the acoustic environments, the ASC framework exploits deep machine learning methods, feature learning with matrix factorization as well as image features from TFR [5], [23]. The deep learning based techniques such as convolutional neural networks (CNN) are popularly used for acoustic scene classification tasks in the DCASE 2016 and DCASE 2017 challenges [23]. The spectrogram and Mel-energy TFR have been widely used as the input to a CNN for audio classification problems [24]. In the DCASE 2016 challenge, Lidy et al. [8], however show that a CQT TFR performs better than the Mel-energy TFR in providing the input to the CNN. The classification accuracy was increased by 4% by using the CQT TFR compared to the Mel-energy TFR. The deep learning performance can vary significantly with different feature representation and architecture and a large amount of data is required to train the feature learners [25]. In [9], [26] and [27], it has been demonstrated that the acoustic scenes can be learned from a TFR using matrix factorization techniques. Matrix factorization methods using Principal Component Analysis (PCA) and Nonnegative Matrix Factorization (NMF) have been explored with different variants and tuning strategies to further improve the classification accuracy. A nonnegative task-driven dictionary learning was ranked among the top performing system in the DCASE 2016 challenge [9].

Also, particular attention has been paid on using computer vision techniques to extract features from time-frequency images of acoustic scenes. Image features such as LBP, Sub-band Power Distribution (SPD) [28] and HOG have been exploited in ASC. The CQT and spectrogram have been widely used as TFRs to integrate with the feature learning method or to be coupled with hand-crafted features such as HOG and LBP [8], [29], [16], [30], [9]. The texture features extracted from the TFRs capture the time and frequency discriminative patterns of the audio structure. Ye et al. [31] incorporated the statistics of local pixel values of the spectrogram into the LBP. In [30], LBP features were extracted from the spectrogram and the bag-of-features model was applied to generate LBP-Codebook features. Recently, [32] proposed the application of LBP to capture the temporal dynamic features of MFCC's representation. Nevertheless, they showed that the application of LBP on MFCC and the complementary spectral features do not provide a significant improvement. The TFR representation retains the important spectral and temporal structure

appropriate for ASC and is more suited to being viewed as 2D texture images compared to MFCC.

In our previous work [16], by using the CQT TFR as a texture pattern, the LBP operator for feature extraction was applied. A TFR zoning mechanism for LBP features provides a simple solution to extract spectrally relevant local features, which better characterize the audio TFRs. On the other hand, [29] exploited HOG features with a pooling strategy on the CQT TFRs to locally analyze the features. Pooling over time, which averages all histograms along the time axis gave better performance than when averaging over the frequency axis. A similar idea using the Mel-spaces zones with the Short Time Fourier Transform (STFT) has been used in [33] for music genre classification and has provided a 3% accuracy improvement compared to the global features. To date, only STFT and CQT TFRs have been used together with standard computer vision based image features like LBP and HOG. In this work we propose the VQT as a better suited TFR for acoustic scenes and identify a variant of LBP which better exploits the characteristics of these TFR images.

III. Proposed Time Frequency Representation with Texture Features

Our proposed framework uses a TFR of the acoustic signals to be coupled with a variant of LBP for feature extraction. The audio signals are transformed to a log-scale frequency band TFR using the VQT and the TFR pixel values are given by the log magnitude of the VQT. Considering the VQT TFR as a texture pattern, we apply an image feature method, AECLBP for feature extraction. Since the acoustic scenes have large variations in time, each TFR is segmented along the frequency axis to extract local information, which better highlight distinctive spectral regions of the scene. The TFRs are divided into n number of zones to obtain the localized information of a given texture pattern. Then, the features are extracted from each zone of the TFR and the histogram of the AECLBP is computed. Subsequently, the features are concatenated to form an enhanced AECLBP feature vector. Furthermore, the concatenated features are fused with HOG and provide complementary features to the AECLBP. Finally, for classification, the SVM classifier is used. Figure 1 illustrates the conceptual framework of the proposed system and the key processing stages are discussed in the following sections. We start by introducing our novel utilization of the VQT time-frequency representations in III.A, followed by the proposed application of the AECLBP feature extraction method (III.B). Then, we explain the zoning mechanism and the feature level fusion in III.C and III.D respectively.

A. Log-Frequency Resolution of Time-Frequency Images

Transforming an acoustic signal to a TFR will portray the patterns of the power spectrum across time and frequency instances. Figure 2 depicts the TFRs for a 'bus' scene. Using the log-scaled STFT spectrogram, shown in Figure 2a, the linear frequency spacing does not emphasise the information at the lower frequencies, hence the need for some form of

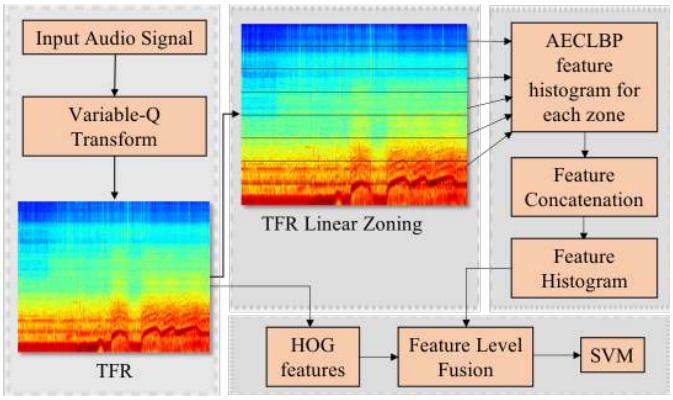


Fig. 1. The conceptual framework of the proposed ASC system

log scale frequency spacing arises. Conversely, the CQT provides a TFR in a log scale frequency resolution that is designed to map the scale of Western music. The CQT TFR has been adapted to the acoustic scene analysis and shows better resolution than the STFT spectrogram as shown in Figure 2b. However, the constant Q-factors of the CQT provides no control over the time resolution, which is poor at the lower frequencies. Whereas by using the VQT representation, shown in Figure 2c, the audio information can be presented clearer even at the lower frequency to preserve the information. The VQT is also computed using the log scale TFR but the Q-factor is allowed to vary in providing a better representation of TFR.

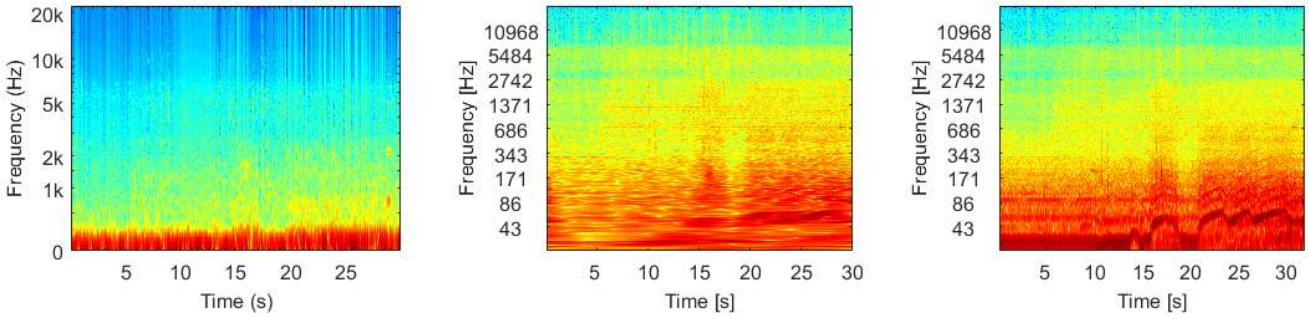
The VQT has smoothly varying Q-factors, which represent the TFR with bandwidths that are constant on the auditory critical-band scale, unlike the commonly used CQT in which Q-factors are constant throughout the log-frequency scale and the analysis window sizes increases towards the lower frequencies. Auditory filters in the human auditory system are approximately constant-Q only for frequencies above 500 Hz but have a constant bandwidth towards lower frequencies [7]. Texture patterns presented by the VQT provide the finer representation of time-frequency information compared to the CQT.

In CQT, the centre frequencies of the frequency bins are geometrically spaced and their Q-factor is constant. The Q-factor, α is given by the ratio of the bandwidth, B_k of the frequency bin, k to the centre frequencies, f_k.

$$\alpha = \frac{B_k}{f_k} \tag{1}$$

However, one considerable problem is the fact that the time-domain frames get very long towards lower frequencies hence decreasing the time resolution at these frequencies. In the regard, a variable-Q representation offers increased temporal resolution at lower frequencies compared to the Constant-Q representation. An additional parameter, γ was introduced by [7] to allow a smooth decrease of the Q-factors of the bins towards low frequencies:

$$B_k = \alpha f_k + \gamma \tag{2}$$



(a) Spectrogram (b) CQT (c) VQT

Fig. 2. An example of the time frequency representation of the 'bus' scene for (a) spectrogram (b) CQT and (c) VQT

In the constant-Q case, the Q-factor, α is a constant when $\gamma = 0$. The comparison of the TFR at different values of γ is shown in Figure 3. It should be noticed that the TFR patterns are similar at the high frequencies, but the VQT shows better temporal resolution at the low frequencies. The γ parameter provides flexible control of temporal resolution at low frequency zone, which is more adaptable to the acoustic scene signal.

B. Adjacent Evaluation Completed LBP Feature Extraction

The LBP feature has attracted increasing interests in the computer vision community [10], [34]. Ojala et al. [35] introduced the LBP with circular neighborhoods allowing any radius and number of neighbors in order to be able to deal with structure at various scales. For texture classification, the pattern value depends on the illumination, rotational variance of the texture and vulnerability to camera capture degradation (blur, pixel noise, etc.). However, in the case of an acoustic signal TFR, the texture pattern depends on the intensity distribution across frequencies and time instances.

The intensity distribution of these time-frequency structures are assumed to be the pixel values in the TFR texture image. The LBP generates a binary code for every pixel of an image. Figure 4 shows an example of a TFR with sample pixel values. The LBP encodes the sign of the relative intensities of a pixel to its neighbor to capture the micro-structures of the TFR texture.

LBP considers each pixel of an image and it is calculated by comparing each central pixel, g_c, with its neighboring pixels, g_p, where the radius, R is the distance between the central pixel and P neighboring pixels. The radius, R, determines the scale of the micro-structures while the number of neighbors, P, captures the pixel information on the image texture. If the value of the neighboring pixel, g_p, is greater than the level of central pixel, g_c, the binary bit is set to 1, otherwise it is set to 0.

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p, \;\; s(x) = \begin{cases} 1 & x \geqslant 0 \\ 0 & x < 0 \end{cases} \tag{3}$$

This traditional LBP is sensitive to the change of pixel values which only considers the sign difference of the energy distribution from the TFR. On the other hand, the AELBP [11] constructs an adjacent evaluation window which is around the neighbor pixel to deal with the neighboring pixel that may vary significantly. The TFR pattern is a random texture and the random changes of the pixel values can be avoided by averaging the neighboring pixels in the evaluation window. An adjacent evaluation window around the neighboring pixel is constructed to modify the threshold scheme of LBP and this reduces the interference from the saltation of neighbors' values to improve noise robustness. The difference between the LBP and the AELBP is that the AELBP replaces the neighboring pixel, g_p, with an evaluation center pixel, a_p, which is the average value of the neighboring pixel in the evaluation window.

$$AELBP_{P,R} = \sum_{p=0}^{P-1} s(a_p - g_c)2^p, \;\; s(x) = \begin{cases} 1 & x \geqslant 0 \\ 0 & x < 0 \end{cases} \tag{4}$$

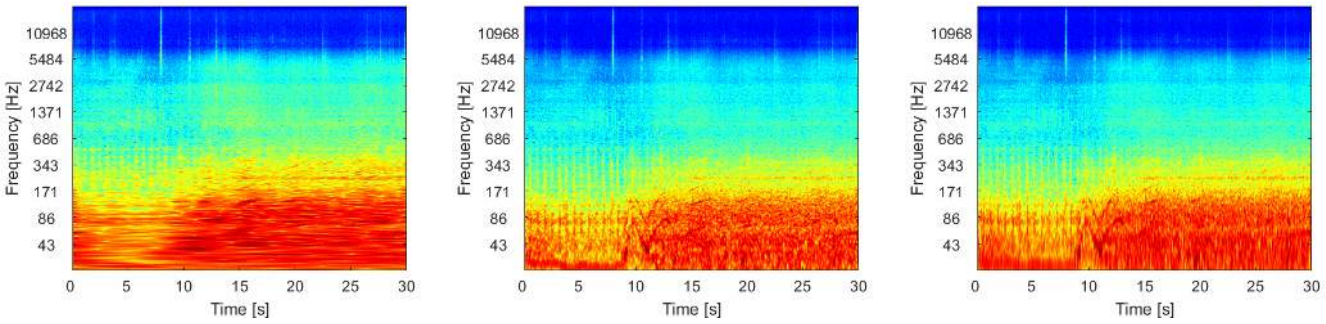
Figure 5 depicts an example operation of AELBP. Adjacent to the center pixel, g_c, an evaluation window of size 3 by 3 is set up around neighboring pixels, g_p. From the evaluation window, the value of evaluation center, a_p is obtained by calculating the average of the pixel values in the evaluation window excluding the neighboring pixel, g_p. The evaluation center, a_p, replaces the traditional neighboring pixel, g_p, value. If the value of the evaluation center pixel, a_p, is greater than the level of central pixel, g_c, the binary bit is set to 1, otherwise it is set to 0. This encoding strategy will make the local binary pattern more stable with the random change of neighboring pixels values, thus reducing the interference of noise.

In order to extract more discriminative patterns, Song et al. [11] integrated the AELBP with CLBP [12] to derive the AECLBP. In AECLBP, the image local differences are decomposed into two complementary components i.e. the sign difference (s_p) and the magnitude difference (m_p) which is described as:

$$s_p = s(a_p - gc) \tag{5}$$

and

$$m_p = |a_p - g_c| \tag{6}$$



(a) $\gamma = 0$ (b) $\gamma = 10$ (c) $\gamma = 20$

Fig. 3. TFRs of a 'car' scene with different values for γ are shown. (a) is for the constant-Q case where $\gamma = 0$, (b) and (c) are for the variable-Q case where $\gamma = 10$ and 20 respectively.

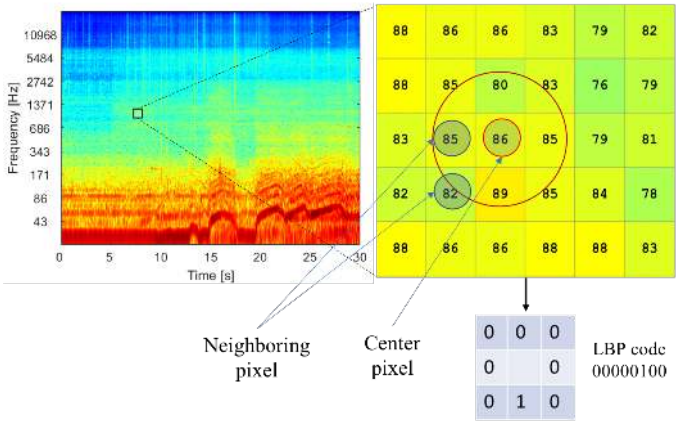


Fig. 4. An illustration of pixel value from the VQT TFR. A LBP encoding is computed by comparing the relative intensities of the pixel values to its neighbors.

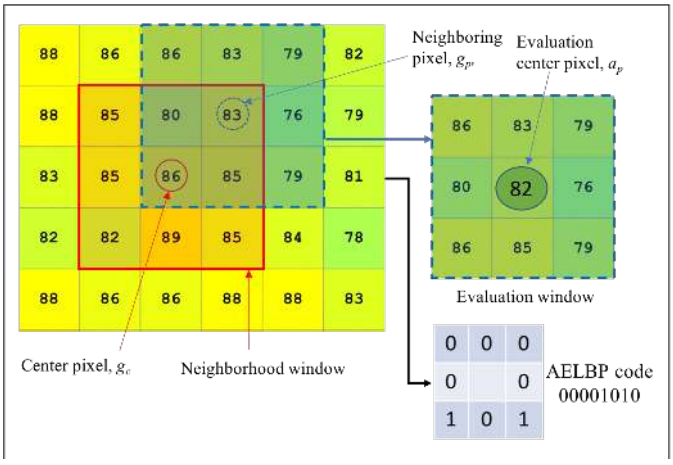


Fig. 5. An example of AELBP operation, where the neighboring pixel values are replaced by the evaluation center pixel.

The sign difference of AECLBP, denoted as AECLBP_S is the same as the AELBP defined in Eq. (4) and the magnitude component denoted as AECLBP_M is defined as in Eq. 7.

$$AECLBP\_M_{P,R} = \sum_{p=0}^{P-1} t(m_p, c)2^p,\ t(x,c) = \begin{cases} 1 & x \geqslant c \\ 0 & x < c \end{cases} \quad (7)$$

The threshold, c, is set as the mean value of the magnitude difference, m_p, calculated over the whole image. However, in our case, the threshold, c, is calculated as the mean value of magnitude difference, m_p, for each zone. Thus, this will make the encoding more localized within each zone. The sign operators, AECLBP_S and magnitude operators, AECLBP_M are jointly combined by simple concatenation as in Eq. 8.

$$AECLBP = [AECLBP\_S \quad AECLBP\_M] \quad (8)$$

The information contained in the magnitudes of local differences is included as complementary features to the sign component of traditional LBP. By using the integrated feature which combines the sign and magnitude components, more discriminative information can be obtained.

C. Global and Local Features

In this experiment, the audio files are not independently processed frame by frame but are considered as a segment of the texture image from the TFR. Analyzing the TFR as a global feature will average out the distinctive regions of the TFR. Each TFR segment contains local time-frequency information of various events occurring in the scene. It is observed that in the acoustic scenes, the time sequence characterizing the same scene can appear in a different order. The frequency signature of the scene is represented by segmenting the time-frequency images along the frequency axis. As proposed in our previous work [16], the TFR image is divided into n horizontal zones of equal size as shown in Figure 6. The zoning size, is dependent on the TFR height, h and the number of zones, n and is given by, $z = round(h/n)$. The number of zones, n, was varied empirically to get the best results and we found $n = 10$ provided the optimal number of zones.

The local AECLBP histograms in each zone are extracted and concatenated to form a final AECLBP feature vector. The local histogram will gather in the features pertinent to the relevant spectral information to discriminate between the acoustic scenes. By zoning the TFR, a unique value for the threshold, c, of Eq. 7 can be derived. It is calculated as the mean value of the magnitude differences, m_p, in each zone to represent the local intensity information. Our novel

enhanced encoding in a local neighborhood by looking at the different threshold values represents local intensity difference information that is not captured by the traditional LBP features.

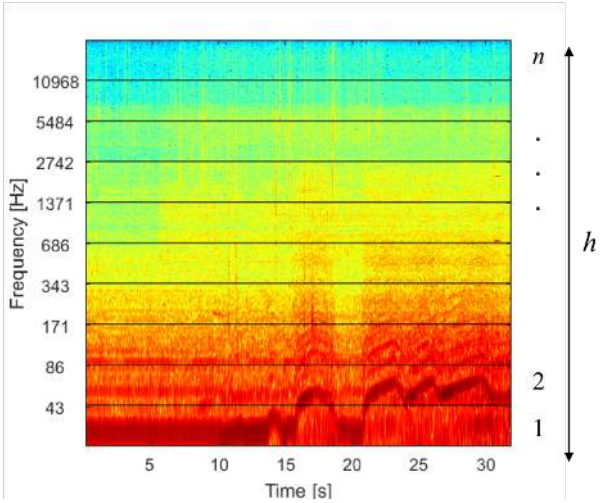


Fig. 6. Uniform zoning of TFR to extract the local features.

D. Fusion

The HOG features emphasize different characteristics of the TFR compared to AECLBP. HOG features derive the distribution of gradients at different orientations which better deals with edges and corners compared to LBP or its variant [36]. In HOG, the gradient of the pixels of the time-frequency images to characterize the spectro-temporal variations of acoustic events occurring in a scene is computed [29]. In order to further improve the recognition rate, we fused the AECLBP and HOG features to complement the different textures of the VQT pattern by concatenating the HOG directly with the AECLBP. We applied early fusion as it operates at the feature level to combine all the feature vectors extracted into a single enhanced vector.

IV. Experiment

A. Dataset and system parameters

Our experiments were performed on the DCASE 2016 dataset [20] for ASC. The DCASE 2016 development dataset contains 1170 audio files of 15 diverse labelled indoor and outdoor scene classes i.e. lakeside beach, bus, cafe/restaurant, car, city center, forest path, grocery store, home, library, metro station, office, urban park, residential area, train and tram. There are 78 samples for each acoustic scene and the duration for each recording is 30 seconds. The development set was further partitioned into four folds of training and testing sets following the set up provided by the DCASE 2016 organizer.

For each fold, the accuracy is calculated for each scene class i, where $i = 1, 2, \ldots, k$ and k is the number of classes. The accuracy is calculated as the number of total correct scenes divided by the total number of test scenes. Accuracies for each fold are obtained by averaging the 15 scene class accuracies. Finally, the overall accuracy is evaluated by averaging the four folds' accuracies.

For the calculation of the VQT, we used the toolbox from [7]. We then enhance the image by resizing the TFR to get an image, which produces less blurring of the edges [37]. Also by resizing, we obtain a uniform TFR that is robust across different audio signals without depending on the sampling rate, signal length and VQT parameters. We have chosen to resize all TFRs to 512 by 512 pixels, which follows the setting from [29] to preserve the time-frequency structures of the audio scene. The configuration parameters for the TFR computation are shown in Table I.

TABLE I
VQT PARAMETER EXPERIMENTAL VALUES

Gamma, γ	20
Bin per Octave, B	96
Maximum frequency, F_{max}	22,050 Hz
Minimum frequency, F_{min}	22.05 Hz
Sampling frequency, F_s	44,100 Hz

For the LBP and AECLBP features, we performed preliminary experiments by varying the values of the radius, R and the number of neighboring pixels P. More information on texture micro-structures can be captured with increased number of P neighbouring pixels, but at the cost of high dimensionality. We found that $P = 12$ and $R = 2$ provided the best performance for the feature descriptor. We used uniform LBP encoding from [35], which best corresponds with the primitive micro-features. The LBP code is considered as a uniform pattern when the bitwise transitions from 0 to 1 or vice versa is at most twice in a single LBP code. The non-uniform patterns are considered as noisy patterns and hence grouped into an extra bin. The uniform patterns with P neighboring pixel yields $P(P-1)+3$ feature dimensions. For example, for $P = 12$, the size of the LBP features is 135 for each zone. Following the concatenation of the features from 10 zones, the size of LBP features is 1350. It should be noted that the feature dimension for AECLBP is 2700, double the number of features for LBP. The HOG features are obtained using the best performing features from [29] and the feature dimension is 1536. For classification, the SVM classifier with linear kernel is implemented using the LIBSVM toolkit [38] and a "one-against-one" approach for multi-class classification was used.

B. Experimental Results and Analysis

In order to evaluate the effect of the proposed input VQT TFR compared to CQT, we first evaluated the performance of the two TFRs i.e. CQT and VQT with LBP features. We performed preliminary experiments by varying the values of the radius, R and neighboring pixel P (designated as LBP(P, R)). Table II shows the average accuracy across the 4-fold cross validation. The results are presented for γ= 20 (VQT case) and γ= 0 (CQT case), which show that the performance with VQT provides a 2.2% improvement over

CQT. The VQT with LBP features improves the classification accuracy of the baseline system from 71.3% to 80.2%. This result clearly demonstrates the benefit of using a VQT representation. The better temporal resolution of the VQT at the lower frequency may explain the better performance of VQT over CQT.

TABLE II
AVERAGE CLASSIFICATION ACCURACY (%) BY USING CQT AND VQT WITH ORIGINAL LBP FEATURES COMPARED TO THE BASELINE SYSTEM.

Method	**Average Accuracy (%)**
Baseline	
MFCC + GMM	**71.3**
CQT, γ = 0	
LBP(8,1) + SVM	**76.0**
LBP(12,2) + SVM	**78.0**
LBP(16,2) + SVM	**79.3**
VQT, γ = 20	
LBP(8,1) + SVM	**75.3**
LBP(12,2) + SVM	**80.2**
LBP(16,2) + SVM	**80.3**

Then, we carried out an analysis using AELBP. The AELBP performance is presented in Table III, which shows an improvement of 2.1% in accuracy compared to the traditional LBP. By utilizing AELBP which used an evaluation window, the spectro-temporal information can be extracted as it considers the influence of other neighboring pixels around the center pixel. The neighboring pixels of the TFRs convey a different intensity that cannot be captured by the traditional LBP since the traditional LBP features are calculated over a local neighborhood. The adjacent evaluation window function overcomes the issue of random intensity of pixel values that might be considered as noise in ASC.

TABLE III
CLASSIFICATION ACCURACY (%) OF USING VQT WITH AELBP FEATURES

VQT, γ = 20	**Average Accuracy (%)**
AELBP(8,1) + SVM	**78.9**
AELBP(12,2) + SVM	**82.3**
AELBP(16,2) + SVM	**82.1**

We examine the effect of using different gamma, γ, for ASC shown in Table IV. We can see from the results, that the gamma value is important in improving the classification accuracy. This confirms that the variable-Q factor is essential to retain the important spectral and temporal structure of the acoustic signal. In our experiment, using gamma, γ = 20 provides the best performance accuracy with the AELBP features.

In order to extract more discriminative patterns of the TFRs, we carried out an analysis using AECLBP, which is reported in Table V. By integrating the VQT with AECLBP, the acoustic scene classification accuracy is further improved. The AECLBP yielded a 3% improvement in accuracy compared to the traditional LBP and up to 11.9% improvement over the baseline system. In AECLBP, we

TABLE IV
CLASSIFICATION ACCURACY (%) OF USING DIFFERENT GAMMA FOR VQT WITH AELBP FEATURES

AELBP (12,2)	**Average Accuracy (%)**
γ = 0	**80.8**
γ = 5	**81.0**
γ = 10	**80.9**
γ = 15	**81.3**
γ = 20	**82.3**

obtained an accuracy of 83.2% compared to 80.2% using traditional LBP. AECLBP incorporates the magnitude component as additional information to extract more discriminative patterns. By considering both sign and magnitude, 2 different codes are assigned to each pixel in the TFR. Joint distribution of these codes resulted in more accurate information of the acoustic scene. Notice that, by using $P = 12$ and $R = 2$ this provided the best performance for the feature descriptor as described in Section IV.A. Also, we compare the the performance of the zoning technique to the global TFR (without zoning). The result shows that the zoning technique provides a better accuracy by an 8.7% improvement compared to the global TFR. This confirms that the zoning technique is beneficial to retain the spectral information pertinent to each zone.

The Student's t-test was employed to evaluate the statistically significant differences between CQT and VQT with the LBP (12,2) features. The statistical test has shown that the p-value of the statistical test was less than the critical value at 85% confidence level ($p < 0.15$). However the proposed system of VQT with AECLBP (12,2) achieves 5.2% improvement over CQT with LBP (12,2), which is competitive with the top performing systems for DCASE2016 challenge.

TABLE V
CLASSIFICATION ACCURACY (%) OF VQT+AECLBP WITHOUT ZONING, VQT+HOG AND VQT+AECLBP

VQT, γ = 20	**Average Accuracy (%)**
AECLBP(12,2) + SVM Without zoning	**74.5**
HOG + SVM	**81.0**
AECLBP(8,1) + SVM	**80.2**
AECLBP(12,2) + SVM	**83.2**
AECLBP(16,2) + SVM	**82.9**

As HOG emphasizes different capabilities in image analysis, we also experimented on the performance of HOG with VQT. Table V presents the performance of VQT with HOG features. We can see that the accuracy score of AECLBP features outperforms HOG features by 2.2%. This make sense as the HOG features are not able to totally capture the fine-grained discriminative features compared to AECLBP.

We investigate the per-class accuracies by having a look at the classification accuracy across the 15 scene classes for the baseline, LBP, HOG and AECLBP features. The obtained result is shown in Figure 7. The results show that AECLBP

and HOG features produced the highest accuracy for the majority number of scene classes. The HOG features performed well for vehicle sounds (car and bus), while AECLBP performed better for outdoor sounds (beach, city centre and forest path). On the other hand, the baseline system exhibited a volatile performance, with the best performance only for indoor scenes such office, café and home but a weak performance for other classes such as park and train. The baseline system is indeed superior for these three classes which are typically mildly noisy indoor scenes. One possible explanation is that the use of extracting visual based features from a TFR loses some critical information which discriminates these scenes. The pattern variations caused by the audio noise may significantly change the TFR representation. However, our system performs better overall and is less volatile compared to the baseline system, but this does highlight further improvements are possible by improving the performance of these three classes. Also, it should be noted that all features had difficulties in discriminating scenes such as park, residential area, train and home with an accuracy of less than 80%.

Further, with the fusion of AECLBP with HOG, the performance accuracy is significantly improved compared to the individual features as shown in the confusion matrix of Figure 8. The classification accuracy of park, residential area and home improve as HOG and AECLBP emphasize different capabilities in image analysis. HOG is excellent at capturing edges and corners in images, while AECLBP captured the fine-grained of local texture pattern. This explains the successful fusion in providing complementary features. However, in some cases e.g., cafe/restaurant the weak performance of HOG provides a lower accuracy than AECLBP alone.

We also compared the accuracy scores of the proposed fusion method with the state-of-the-art [32], [8] and the top non-neural network based system from DCASE 2016 [9], as shown in Table VI. Compared to [32], which adopted LBP features based on MFCC TFRs, and [8] that uses CQT with CNN, our use of the VQT with AECLBP and HOG yielded a superior classification accuracy. The VQT time-frequency representation retains the important spectral and temporal structure of the acoustic signal and is more suited to being viewed as texture images compared to MFCC. Our proposed method obtained an accuracy of 85.5% to outperform one of the top systems [9] using CQT with matrix factorization. This confirms that vision based features extracted from TFR images provides competitive performance in line with deep learning and matrix factorization methods.

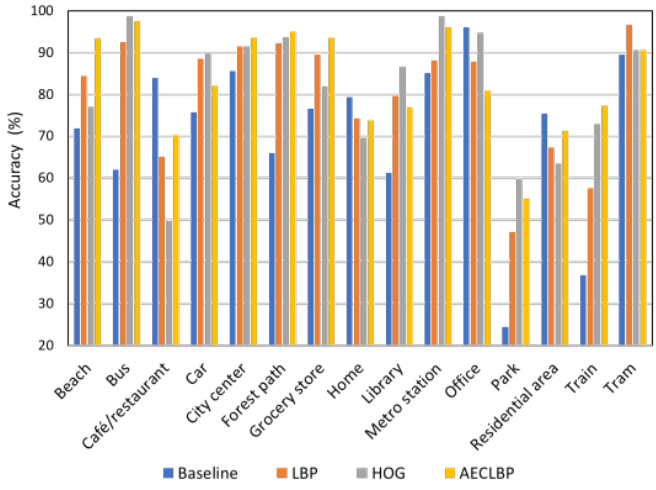


Fig. 7. Performance comparison of different features on the DCASE 2016 dataset.

TABLE VI
ACCURACY SCORES OF THE PROPOSED FUSION FEATURES COMPARED TO THE STATE-OF-THE-ART RESULTS.

Method	Accuracy (%)
LBP+MFCC[32]	80.3
CQT + CNN[8]	80.3
CQT+Matrix Factorization[9]	83.8
VQT+AECLBP+HOG	**85.5**

V. CONCLUSIONS

This work has demonstrated the capability of AECLBP features extracted from a time-frequency representation for acoustic scene classification. The micro structure of the image texture is different from the TFR, hence the variant of LBP that is adaptable to the acoustic signal representation and better suited for ASC has been presented. The VQT TFR retains the important spectral and temporal structure of the acoustic signal and is suitable to be viewed as texture images for ASC. VQT is more adaptable to the acoustic signal since it enhances the TFR resolution representation. The unique combination of VQT and AECLBP provides a better discriminative performance over the CQT and LBP with a 5.2% improvement in classification accuracy. The proposed uniform zoning with AECLBP allows different threshold values to be calculated relevant for each zone. Furthermore, the fusion of AECLBP with HOG features further improves the results providing a state of the art performance. However, due to the similarity characteristics of the acoustic scene classes, the VQT and AECLBP parameters could be customized for specific scene classes in order to offer local contrast and variance information, which is necessary for ASC. Our work shows the potential of image-based variants, with deep learning networks in future work for ASC.

REFERENCES

[1] R. Radhakrishnan, A. Divakaran, and P. Smaragdis, "Audio analysis for surveillance applications," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.*, pp. 158–161, 2005.

[2] A. Härmä, M. F. Mckinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," *IEEE International Conference on Multimedia and Expo*, no. 1, pp. 634–637, 2005.

[3] R. G. Malkin and A. Waibel, "Classifying user environment for mobile applications using linear autoencoding of ambient audio," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 509–512, 2005.

[4] D. Battaglino, A. Mesaros, L. Lepauloux, L. Pilati, and N. Evans, "Acoustic context recognition for mobile devices using a reduced complexity SVM," *European Signal Processing Conference*, pp. 534–538, 2015.

	Beach	Bus	Café/restaurant	Car	City center	Forest path	Grocery store	Home	Library	Metro station	Office	Park	Residential area	Train	Tram	Class-wise classification accuracy (%)
Beach	**69**	0	0	0	3	2	0	0	0	0	0	3	1	0	0	88
Bus	0	**75**	1	0	1	0	0	0	0	0	0	0	0	0	1	96
Café/restaurant	0	1	**48**	0	0	0	20	0	0	6	0	0	1	2	0	62
Car	1	3	0	**73**	0	0	0	0	0	0	0	0	0	1	0	94
City center	0	0	0	0	**74**	0	0	0	0	0	0	0	4	0	0	95
Forest path	1	0	0	0	0	**76**	0	0	0	0	0	0	1	0	0	97
Grocery store	0	0	3	0	0	0	**74**	0	1	0	0	0	0	0	0	95
Home	0	0	2	0	0	0	2	**61**	9	0	4	0	0	0	0	78
Library	0	1	3	0	0	0	0	0	**66**	1	1	0	0	5	1	85
Metro station	0	0	0	0	0	0	1	0	0	**76**	0	1	0	0	0	97
Office	0	0	0	0	0	0	0	3	2	0	**72**	0	0	1	0	92
Park	0	0	0	0	5	14	0	0	4	0	0	**50**	5	0	0	64
Residential area	0	1	0	0	4	2	0	0	0	0	0	13	**58**	0	0	74
Train	1	7	3	1	0	0	0	5	0	2	1	0	1	**57**	0	73
Tram	0	0	0	0	1	0	2	4	0	0	0	0	0	0	**71**	91

Fig. 8. The confusion matrix obtained on the development set by the fusion of AECLBP and HOG features. The rows correspond to the true labels and the columns to the predicted labels.

[5] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[6] C. Schörkhuber and A. Klapuri, "Constant-Q Transform Toolbox for Music Processing," in *7th Sound and Music Computing Conference*, 2010, pp. 3–64.

[7] C. Schörkhuber, A. Klapuri, N. Holighaus, and D. Monika, "A Matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution," *AES International Conference on Semantic Audio*, pp. 1–8, 2014.

[8] T. Lidy and A. Schindler, "CQT-based convolutional neural networks for audio scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, September 2016, pp. 60–64.

[9] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Feature learning with matrix factorization applied to acoustic scene classification," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 6, pp. 1216–1229, 2017.

[10] L. Liu, P. Fieguth, Y. Guo, X. Wang, and M. Pietikäinen, "Local binary features for texture classification: Taxonomy and experimental study," *Pattern Recognition*, vol. 62, pp. 135–160, 2017.

[11] K. Song, Y. Yan, Y. Zhao, and C. Liu, "Adjacent evaluation of local binary pattern for texture classification," *Journal of Visual Communication and Image Representation*, vol. 33, pp. 323 – 339, 2015.

[12] Z. Guo, L. Zhang, and D. Zhang, "A completed modeling of local binary pattern operator for texture classification," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1657–1663, 2010.

[13] Y. Zhao, D. S. Huang, and W. Jia, "Completed local binary count for rotation invariant texture classification," *IEEE Transactions on Image Processing*, vol. 21, no. 10, pp. 4492–4497, Oct 2012.

[14] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1635–1650, June 2010.

[15] T. Komatsu, T. Toizumi, R. Kondo, and Y. Senda, "Acoustic event detection method using semi-supervised non-negative matrix factorization with a mixture of local dictionaries," *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, pp. 45–49, September 2016.

[16] S. Abidin, R. Togneri, and F. Sohel, "Enhanced lbp texture features from time frequency representations for acoustic scene classification," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 626–630, March 2017.

[17] N. Sawhney and P. Maes, "Situational awareness from environmental sounds," *Project Rep. for Pattie Maes*, pp. 1–7, 1997.

[18] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 321–329, 2006.

[19] S. Chu, S. Narayanan, and C.-C. Kuo, "Environmental sound recognition with time- frequency audio features," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 6, pp. 1142–1158, 2009.

[20] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," *European Signal Processing Conference*, pp. 1128–1132, 2016.

[21] D. Giannoulis, D. Stowell, E. Benetos, M. Rossignol, and M. Lagrange, "A database and challenge for acoustic scene classification and event detection," *European Signal Processing Conference*, pp. 1–5, 2013.

[22] G. Roma, W. Nogueira, and P. Herrera, "Recurrence Quantification Analysis features for environmental sound recognition," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 0–3, 2013.

[23] *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*. Tampere University of Technology. Department of Signal Processing, 2016.

[24] A. Mesaros, T. K. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 26, no. 2, pp. 379–393, 2017.

[25] J. Li, W. Dai, F. Metze, S. Qu, and S. Das, "A comparison of deep learning methods for environmental sound detection," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 126–130, March 2017.

[26] A. Rakotomamonjy, "Supervised representation learning for audio scene classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1253–1265, 2017.

[27] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Acoustic scene classification with matrix factorization for unsupervised feature learning," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6445–6449, 2016.

[28] V. Bisot, S. Essid, and G. Richard, "HOG and subband power distribution image features for acoustic scene classification," *European Signal Processing Conference*, pp. 719–723, 2015.

[29] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time-frequency representations for audio scene classification," *IEEE/ACM*

Transactions on Audio, Speech and Language Processing, vol. 23, no. 1, pp. 142–153, 2015.
[30] D. Battaglino, L. Lepauloux, L. Pilati, and N. Evans, "Acoustic context recognition using local binary pattern codebooks," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2015.
[31] T. K. Ye and Jiaxing, "Acoustic feature extraction by statistics based local binary pattern for environmental sound classification," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3076–3080, 2014.
[32] W. Yang and S. Krishnan, "Combining temporal features by local binary pattern for acoustic scene classification," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 6, pp. 1315–1321, 2017.
[33] Y. M. G. Costa, L. S. Oliveira, A. L. Koerich, F. Gouyon, and J. G. Martins, "Music genre classification using LBP textural features," *Signal Processing*, vol. 92, no. 11, pp. 2723–2737, 2012.
[34] M. Pietikäinen and G. Zhao, "Two decades of local binary patterns: A survey," in *Advances in Independent Component Analysis and Learning Machines*. Elsevier, 2015, pp. 175–210.
[35] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.
[36] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," *IEEE 12th International Conference on Computer Vision*, pp. 32–39, 2009.
[37] D. Han, "Comparison of commonly used image interpolation methods," *Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013)*, pp. 1556–1559, 2013.
[38] C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.

Shamsiah Abidin Shamsiah Abidin received BEng (Hons) Electronics Engineering from University of Multimedia, Cyberjaya, Malaysia in 2003 and MEng Telecommunication Engineering from Royal Melbourne Institute of Technology (RMIT) University, Melbourne, Australia in 2010. She is currently a Ph.D student at the School of Electrical, Electronic and Computer Engineering at The University of Western Australia. Her research interests include time-frequency analysis of audio signal, feature extraction and machine learning applied to acoustic scene analysis.

Roberto Togneri Roberto Togneri (M89-SM04) received the Ph.D degree in 1989 from the University of Western Australia. He joined the School of Electrical, Electronic and Computer Engineering at The University of Western Australia in 1988, where he is now currently an Associate Professor. He leads the Signal Processing and Recognition Lab and his research activities in signal processing and pattern recognition include: feature extraction and enhancement of audio signals, statistical and neural network models for speech and speaker recognition, audio-visual recognition and biometrics, and related aspects of language modelling and understanding. He has published over 150 refereed journal and conference papers in the areas of signal processing and recognition, the chief investigator on three Australian Research Council Discovery Project research grants since 2009, was an Associate Editor for IEEE Signal Processing Magazine Lecture Notes and IEEE Transactions on Speech, Audio and Language Processing from 2012 to 2016 and is currently the Area Editor for the IEEE Signal Processing Magazine, Columns and Forums.

Ferdous Sohel Ferdous A Sohel received PhD degree from Monash University, Australia in 2009. He is currently a Senior Lecturer in Information Technology at Murdoch University, Australia. Prior to joining Murdoch University, he was a Research Assistant Professor/ Research Fellow at the School of Computer Science and Software Engineering, The University of Western Australia from January 2008 to mid-2015. His research interests include computer vision, image processing, pattern recognition, multimodal biometrics, scene understanding, robotics, and video coding. He is a recipient of prestigious Discovery Early Career Research Award (DECRA) funded by the Australian Research Council. He is also a recipient of the Early Career Investigators award (UWA) and the best PhD thesis medal form Monash University. He is a Member of Australian Computer Society and a Senior Member of the IEEE.