IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 23, 2020, accepted September 24, 2020, date of publication September 28, 2020, date of current version October 8, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3027474

An Audio Data Representation for Traffic Acoustic Scene Recognition

DAZHI JIANG[1,2], DONGMIN HUANG[1], YOUYI SONG[3], KAICHAO WU[1], HUAKANG LU[1], QUANQUAN LIU[1], AND TENG ZHOU[1,2,3]

[1]Department of Computer Science, College of Engineering, Shantou University, Shantou 515063, China
[2]Key Laboratory of Intelligent Manufacturing Technology, Ministry of Education, Shantou University, Shantou 515063, China
[3]Center for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong

Corresponding author: Teng Zhou (zhouteng@stu.edu.cn)

This work was supported in part by the NSFC under Grant 61902232 and Grant 61902231; in part by the Natural Science Foundation of Guangdong Province under Grant 2018A030313291 and Grant 2019A1515010943; in part by the Department of Education of Guangdong Province under Grant 2018GXJK048, Grant 2018KZDXM035, and Grant 2017KCXTD015; in part by the Basic and Applied Basic Research of Colleges and Universities in Guangdong Province (Special Projects in Artificial Intelligence) under Grant 2019KZDZX1030; in part by the Shantou University (STU) Scientific Research Foundation for Talents under Grant NTF18006; and in part by the 2020 Li Ka Shing Foundation Cross-Disciplinary Research Grant 2020LKSFG05D and Grant 2020LKSFG04D.

ABSTRACT Acoustic scene recognition (ASR), recognizing acoustic environments given an audio recording of the scene, has a wide range of applications, e.g. robotic navigation and audio forensic. However, ASR remains challenging mainly due to the difficulty of representing audio data. In this article, we focus on traffic acoustic data. Traffic acoustic sense recognition provides complementary information to visual information of the scene; for example, it can be used to verify the visual perception result. The acoustic analysis and recognition, in consideration of its simple and convenient, can effectively enhance the perception ability which only applies visual information. We propose an audio data representation method to improve the traffic acoustic scene recognition accuracy. The proposed method employs the constant Q transform (CQT) and histogram of gradient (HOG) to transfer the one-dimensional audio signals into a time-frequency representation. We also propose two data representation mechanisms, called global and local feature selections, in order to select features that are able to describe the shape of time-frequency structures. We finally exploit the least absolute shrinkage and selection operator (LASSO) technique to further improve the recognition accuracy, by further selecting the most representative information for the recognition. We implemented extensive experiments, and the results show that the proposed method is effective, significantly outperforming the state-of-the-art methods.

INDEX TERMS Feature extraction, acoustic scene recognition, transportation, acoustic material.

I. INTRODUCTION

Traffic acoustic scene recognition (TASR) is a fundamental task, which increases the awareness capabilities of the driving circumstances [1]. TASR may serve as a promising complementary technique for designing safe and reliable automatic driving systems [2], especially when the visual data is temporarily unavailable at some blind spots. Although the acoustic scene recognition task has received many researchers' attention and there are a few successful real-world applications, e.g. robotic navigation [3], and audio forensics [4] existing ASR methods suffer from insufficient recognition accuracy [5]. The TASR is more challenging, since the road scenarios are complex, highly variable, and uncontrollable.

The reason behind their limitation might be that they represent audio data not so informative as to produce a relatively high recognition accuracy [6]. Existing studies represent audio data mainly according to two criteria: time and frequency. Time-based methods [4], [7], [8] represent audio data mainly by waveform analysis, linear prediction coefficient, zero-crossing rate, and the spectral centroid, etc. Frequency-based methods represent audio data by integrating the magnitude spectrum or the power spectrum over the specified frequency bands [9]–[12]. The resulting coefficients of these algorithms can measure the amount of energy present within different sub-bands and can also be expressed as a ratio between the sub-band energy and the total energy to

The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney.

outstand the most prominent frequency regions in the signal. The latest techniques in this line of study try to analyze the nonlinear signals in order to mimic the human perceptual system response [13]–[18].

However, it is not straightforward how to represent audio data, and hence recently, instead of further developing the aforementioned methods, there are a few studies trying to leverage unsupervised learning techniques to represent acoustic data by learning from the distribution of data. For example, Moragues *et al.* [19] adopt the sparse restriction Boltzman machine (SRBM) to learn the audio representation by mel-frequency cepstral coefficients (MFCCs). Such SRBM is a neural network that can learn basic maps from input data, which are similar to those constructed by visual receptors in the human brain [20]. For the problem of acoustic scene recognition, the SRBM could adaptively refine the basic attributes of the signal spectrum, and its activation function could be used to determine the period in which the important acoustics contained. Although learning audio representation directly from the data is theoretically promising, in practice, a method that is able to accurately incorporate the domain knowledge into data representation often yields a far higher classification performance, especially for the acoustic scene recognition task. For example, Rakotomamonjy and Gasso [6] proposed to incorporate the domain knowledge by employing constant Q transformation (CQT) and histogram of oriented gradient (HOG) for the audio data representation. However, such representation was motivated by the characteristics of image processing. Their feature extraction algorithm for acoustic scene recognition includes the following operations. Firstly, they use a time-frequency representation to process the audio signals corresponding to training sets. Then, they represent the frequency by the logarithmic interval frequency. Secondly, by interpolating neighboring time-frequency bins, the constant Q representations can be converted to a gray image of 512 × 512 pixels. Finally, the features are extracted from HOG images by calculating the matrix of the local gradient histogram. The method proposed above is a new idea and opens new horizons for acoustic scene recognition. Although this method substantially impacts the research field, their performance remains to be improved, mainly because of the implementation difficulty of directly operating the HOG algorithm on audio data due to the dimension issue (please refer to [6] for more details). As a result, this method cannot describe the slight fluctuations and also ignores much necessary information for the recognition task.

In order to overcome such limitation, Ye *et al.* [21] suggested increasing the dimension of the features vector. Dessein *et al.* [22] also demonstrated that a combination of employing CQT and Nonnegative Matrix Factorization (NMF) is able to further improve the recognition accuracy on [6]. Similarly, Bisot *et al.* [3] verified that by combining HOG and sub-band power distribution (SPD) techniques is also effective. Although these method consider the subtle cues in the fluctuations, they subsequently increase the dimension of the feature vectors. For the sample processing, Phan *et al.* [23] proposed a sliding window method to solve the high dimensional problem of samples. For an audio sample of a 30s' fragment, a sliding window with a frame length of 500 ms is sheathed to process the sample with the moving of the step length 250 ms. The frame length of 500 ms has 50% overlaps. This method takes every frame as a sample instead of the entire 30s' segment as a sample. This method has been reported that achieved better results than the original HOG method, but obtained weak performance for different audio scenes with some same scene sound effects (such as cafe and quiet street that have the same silent snippets).

In this article, we present a novel and effective method to represent the audio data for TASR, by capturing the global and local expressions of HOG. The proposed method is able to describe fluctuations of HOG and capture more necessary information for recognition, which both mechanisms significantly enhance the classification accuracy. In particular, the characteristics of data fluctuations are extracted on the timeline, and hence the fused features can better represent the HOG descriptor, leading to higher recognition accuracy. Furthermore, we develop a feature selection algorithm based on the least absolute shrinkage and selection operator (LASSO) [24] technique to select representative characteristics from the high-dimensional original feature space, which effectively addressed the limitation of the seminal method [6]. We implemented extensive experiments, and the experimental results show that the proposed method is effective, significantly outperforming the state-of-the-art methods.

The remainder of this article is organized as follows. We detail the global and local representation of the traffic acoustic scenes in Section II. In Section III, we first propose a dataset that contain a large number of traffic acoustic scenes, and conduct extensive experiments to evaluate the recognition performance of the acoustic scene representation. Finally, the concluding section draws some general implications and points to what remains.

II. METHODOLOGY

Fig. 1 illustrates the overview of the proposed framework, which takes an audio signal as the input and outputs the recognition results. Given an audio signal, the CQT is first adopted to convert the audio signal to a CQT spectrogram. Then, a series of HOG descriptors are defined to describe the variations of cumulative oriented gradients in both dimensions for all local regions in the CQT spectrogram. To this end, both local and global features are extracted to enrich the spectral-temporal representation. After that, a large number of high dimensional features are extracted from the CQT spectrogram. Thus, we employ a feature selection algorithm based on the least absolute shrinkage and selection operator (LASSO) [25] to selectively extract the most representative features. Finally, the TASR can be detected by integrating the recognition results by different models.

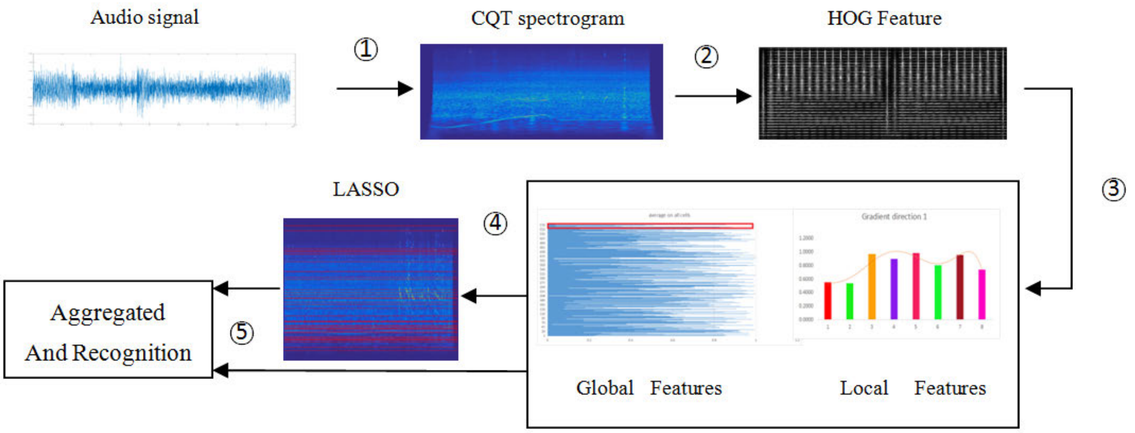


FIGURE 1. The schematic illustration of the TASR framework. An audio signal is first converted to a CQT spectrogram. The HOG features are extracted from the CQT spectrogram. Then, a global feature extraction method and two local feature extraction methods are proposed to extract features in both the time domain and frequency domain. A LASSO is involved for feature selection and reduction.

A. FROM AUDIO TO CQT SPECTROGRAM

We introduce the constant Q transform for traffic acoustic scenes is motivated by the human auditory system. As the output of the transform is effectively amplitude/phase against log frequency, fewer frequency bins are required to cover a given range effectively, and this proves useful where frequencies span several octaves. As the range of human hearing covers approximately ten octaves from 20 Hz to around 20 kHz, this reduction in output data is significant. Since the acoustic scene is difficult, if not impossible to classify in the time domain, we convert the audio clips from the time domain to the frequency domain by the constant Q transformation. The constant Q transformation (CQT) converts a data series into the frequency domain [26]. The CQT is similar to the Fourier transform [27] and the complex morlet wavelet transform. For the acoustic signals in time domain, we define a quality factor Q. The process of this transform can be thought of as a series of algorithmic spaced filters f_k. The kth filter has a spectral width δf_k that equal to the multiple of the previous filter's width:

$$\delta f_k = 2^{\frac{1}{n}} \delta f_{k-1} = 2^{\frac{k}{n}} \delta f_{\min}, \tag{1}$$

where δf_k is the bandwidth of the kth filter, f_k is the central frequency of the lowest filter, and n is the number of filters per octave. The transform exhibits a reduction in frequency resolution with higher frequency bins, which is desirable for the traffic acoustic scenes. The transform mirrors the human auditory system, whereby at lower-frequencies spectral resolution is better, whereas temporal resolution improves at higher frequencies. In another word, For the low-frequency waves, bandwidth will be small, meanwhile a higher frequency resolution is used to decompose similar notes. For the high-frequency waves, bandwidth will be larger, meanwhile a higher temporal resolution is used to track rapidly changing overtones. In this regard, CQT is suitable to represent the acoustic signal in noisy traffic scenes.

B. CQT SPECTROGRAM AND HOG DESCRIPTOR

We scale the CQT spectrogram to 512 × 512 pixels after obtaining the CQT spectrogram by the CQT algorithm. Although the resolution is not required to be the same as Rakotomamonjy [6], we aim to demonstrate the outperformance is obtained by the improvement of the proposed method, not the change of the resolution of the CQT spectrogram. Then, we extract the histogram of oriented gradients (HOG) features from the rescaled CQT spectrogram to analyze the direction information of the local HOG descriptors in time-frequency representation (TFR). The HOG describes the pedestrian characteristics by the gradient histogram. First, the HOG is computed in each small local area. Then, the composition of multiple cells connects histograms together to form a local HOG descriptor. The local HOG descriptors are denoted as feature vectors. The feature vectors are flattened to a final feature vector. The dimension of this final vector depends on the number of bins in the histogram. In this article, the interference factors, such as illumination, are not involved, so the normalization of the input image or the color space is waived. In addition, by adjusting the number of bins, or gradient directions, we can construct a series of HOG feature descriptors correspondingly.

C. GLOBAL AND LOCAL FEATURE EXTRACTION

The flowchart for feature extraction is detailed in Fig. 2. In Fig. 2, the first unit is the original CQT spectrogram. The following five parts illustrate the feature extraction procedure. Firstly, we transform the CQT spectrogram into HOG descriptors, e.g., from unit ① to unit ② in Fig. 2. Next, we transform the HOG descriptors to a list of statistical histograms. To clarify, we take the cell at the bottom right corner of the unit ① as an example. The cell of size 8 × 8 is transformed into a list of HOG vectors in the bottom row of unit ②, which are zoomed in unit ③. We find that the gradient in the same direction is similar to the same class, but varies for different classes. In this regard, we take the corresponding

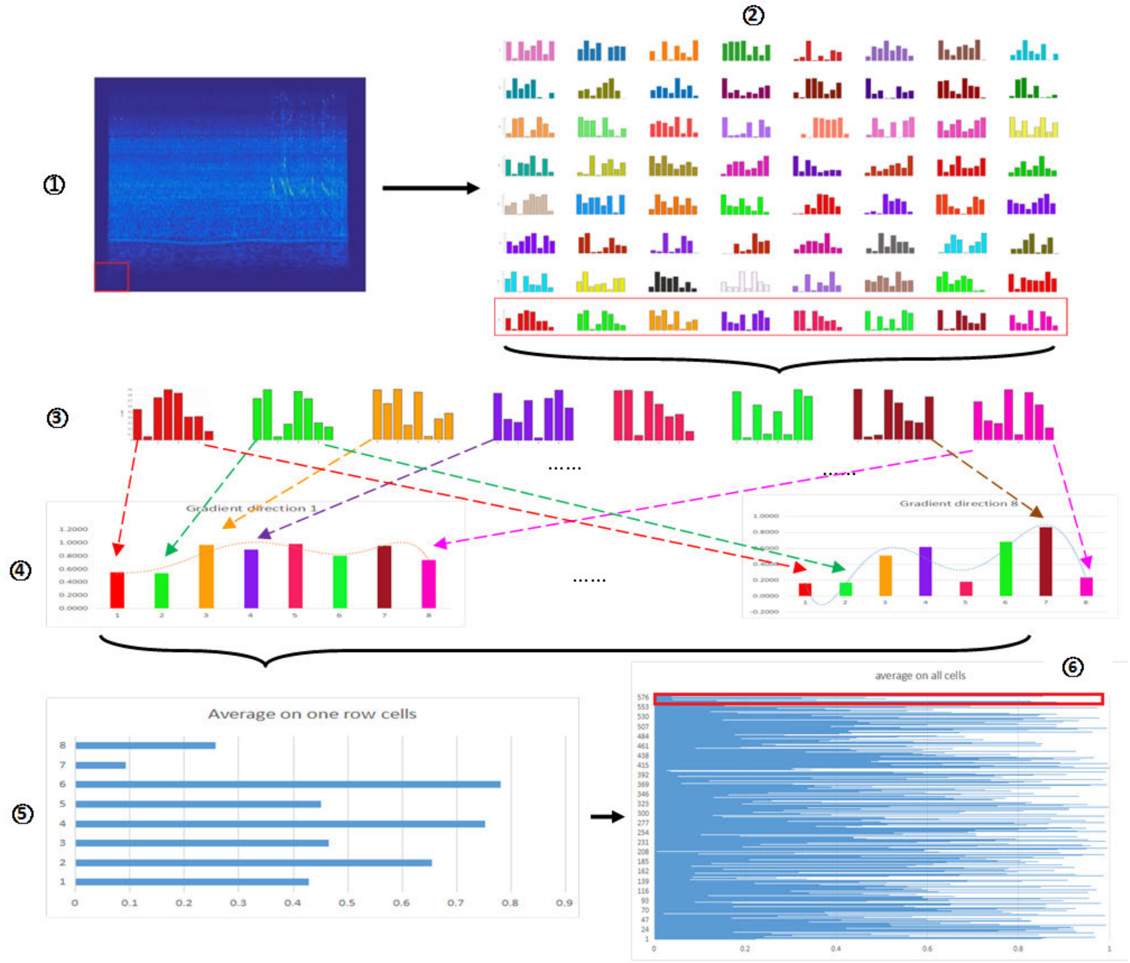


FIGURE 2. **The flowchart of feature extraction. The original CQT spectrogram is divided into 64 cells. The HOG descriptors are extracted from each cell. The gradient of the corresponding direction in each HOG descriptor are aggregated to construct gradient distributions. Then, the gradient distributions are averaged and concatenated for a global feature of the acoustic scene.**

gradient values to reform distribution diagrams. For example, we take the first value of the statistical histogram in each cell from the unit ③ to reform the first distribution diagram named gradient direction 1. This operation results 8 distribution diagrams in unit ④. Then, we average the distributions in each diagram in unit ④. The average results are illustrated in unit ⑤. As a result, the number of cells is changed from 64 into 64, and each cell has 8 statistical components. We assemble the average results into a global feature extraction (GFE) as shown in unit ⑥. Thus, the final dimension of the GFE vector is 64 × 8 = 512.

The GFE reduces the dimensions of the HOG features by performing average pooling along time, and achieves impressive performance. However, it cannot capture a magnitude change along the timeline. In order to address this issue, we also embed a local feature extraction (LFE) method into our model. The CQT spectrogram with 512 × 512 pixels is divided into 64 × 64 cells. Thus, each cell has 8 × 8 pixels. Each cell has 8 gradients, e.g. $c = [f_1, \ldots, f_8]$. We denote the feature vector of cell in the tth row and sth column as g_{ts}. Then, we average the maximum gradient in each cell in the tth row to describe the magnitude change on the time domain.

$$g_t = \frac{1}{64}\sum_{t=1}^{64} \max_{r=1}^{8}(f_r). \tag{2}$$

In this way, the maximum gradient along the timeline is well described. Furthermore, we also present a complementary to record the direction of the maximum gradient as an extended version of the LFE, named LFEx. Similar to Eq. 2, the LFEx averages the index of the maximum gradient as Eq. 3. Then, the LFEx feature is a concatenation of g_t and $\overline{g}_t$.

$$\overline{g}_t = \frac{1}{64}\sum_{t=1}^{64} arg \max_{r=1}^{8}(f_r). \tag{3}$$

D. FEATURE SELECTION AND CLASSIFICATION

After computed the HOG in the spectrogram, the representation composed of histograms is constructed for all cells. The higher resolution of the CQT spectrogram will retain more subtle cues for the subsequent processing. However, if we concatenate all these histograms to yield a final feature vector, the dimension of the final feature vector will be too large. Furthermore, although the details are well retained, the features extracted from the CQT spectrogram may contain irrelevant features and even noises. For example, the total number of cells is 642. Each cell has 8 × 8 pixels and 8 gradient orientations. This results a 642 × 8 = 32768 dimensional feature vector of dimension. The dimension will be further increased, if we reduce the size of the cells or increase the number of orientations when calculating the histogram. Moreover, the noises inside the features may degrade the performance of the recognition results. In this regard, it is essential to reduce the dimension of the final feature vector to eliminate the unrelated features. Least absolute shrinkage and selection operator (LASSO) is a popular tool for sparse linear regression, and it improves the prediction accuracy and interpretability for statistical models by performing feature selection and regularization simultaneously [24]. The key idea of LASSO is to minimize the residual sum of the squares by subjecting the sum of the absolute value of the coefficients to be less than a certain constant. Because of such constraint, the LASSO tends to produce zero coefficients that retain good features during the subset selection and the ridge regression. The ith feature extracted are denoted as $x_j^{(i)}$. The weighting parameters is denoted as $\beta = (\beta_0, \ldots, \beta_p)^\top$. Then, we employ the LASSO to estimate the weighting coefficients of β.

$$\hat{\beta} = arg \min_{\beta} \left[\sum_{i=1}^{N} (y^{(i)} - \beta_0 - \sum_{j=1}^{p} \beta_j x_j^{(i)})^2 \right], \quad \text{subject to} \sum_{j=1}^{p} |\beta_j| \leq t. \tag{4}$$

where N is the number of samples, and p is the dimension of the feature. The parameter $t > 0$ controls the amount of shrinkage that is applied to the estimation of the weighting parameters. Thus, t plays a key role to shrink the weighting coefficients β_j by forcing a part of the elements in β to be 0. Such an operation not only helps in reducing overfitting, but selects discriminated features. Instead of setting a *hard* hyperparameter t, an L_1 penalized simplifies the cost function.

$$\hat{\beta} = arg \min_{\beta} \left[\sum_{i} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j^{(i)})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right]. \tag{5}$$

By increasing the *soft* hyperparameter α, the regularization strength is increased, and the weights are shrunk.

In order to evaluate the discrimination of the extracted acoustic features, we would like to employ classifiers as simple as possible. To achieve this, we employ two simple but frequently used support vector machines, e.g. support vector machine classifier (SVC) with Gaussian kernel and linear support vector machine classifier (linearSVC) for this recognition task. The support vector machine classifier with Gaussian kernel is implemented the same as [28]. The linearSVC from the sklearn is similar to the support vector machine with a linear kernel, but it is more flexible to choose penalties and loss functions, and performs better with a large number of samples.

TABLE 1. The number of samples for each type of vehicles in the traffic acoustic scene dataset.

Vehicle type	Number of samples
Ambulance	60
Bicycle	57
Bus	60
Car	60
Motorbike	58
Police wagon	60
Pumper	60
Train	61
Truck	58
Tube	63

III. EXPERIMENTS

A. DATA DESCRIPTION

We collect 10 classes of traffic acoustic scenes at different locations to construct a traffic acoustic scene dataset to evaluate the performance of the feature extraction method. The data were collected by a recorder and mixer equipped with a Qualcomm Aqstic audio codec (WCD 9335). The raw traffic acoustic slices were collected from 1 minute to 5 minutes at a sampling frequency of 44.1 kHz. The recorded files were saved as AVI format. We cut the raw acoustic slices were cut into 30 seconds per clips. The number of samples in each class is listed in Table 1.

For the clear explanation, we selected 9 CQT spectrograms from 3 typical acoustic scenes e.g. bus, car, and motorbike in Fig. 3. Each subfigure in Fig. 3 is the CQT spectrogram of an acoustic scene. For instance, in the bus CQT spectrograms, the low-frequency line is the acceleration and deceleration of the buses. In the scenes of motorcycle, the higher frequency presented by deeper color is the sound of motorbike engines.

B. DESIGN DECISION

We employ a Matlab toolbox for the constant Q transformation. Then, the CQT spectrogram is scaled to 512 × 512. The VLFeat toolkit is used for HOG conversion. Two different sizes of cell are performed, e.g. 16 × 16 and 8 × 8 on the CQT spectrogram. As a result, the dimension of the feature vector is 64 × 64 × 8 = 32768 for one sample of the traffic acoustic scenes. We illustrated the HOG features of the CQT spectrogram in Fig. 4. The cells in second and third rows of Fig. 4 are 8 × 8. The number of directions for the gradient of each cell is 8 for the second row, and 32 for the third row, respectively. The cells in fourth and fifth rows of Fig. 4 are 16 × 16. The number of directions for the gradient of each cell is 8 for the fourth row, and 32 for the fifth row, respectively. From Fig. 4, we can see the HOG correctly captures the direction of the power spectrum along the high-energy sharp signal comparing with the raw CQT spectrogram. However, it is difficult to choose the right size for the cells and orientations. By performing preliminary

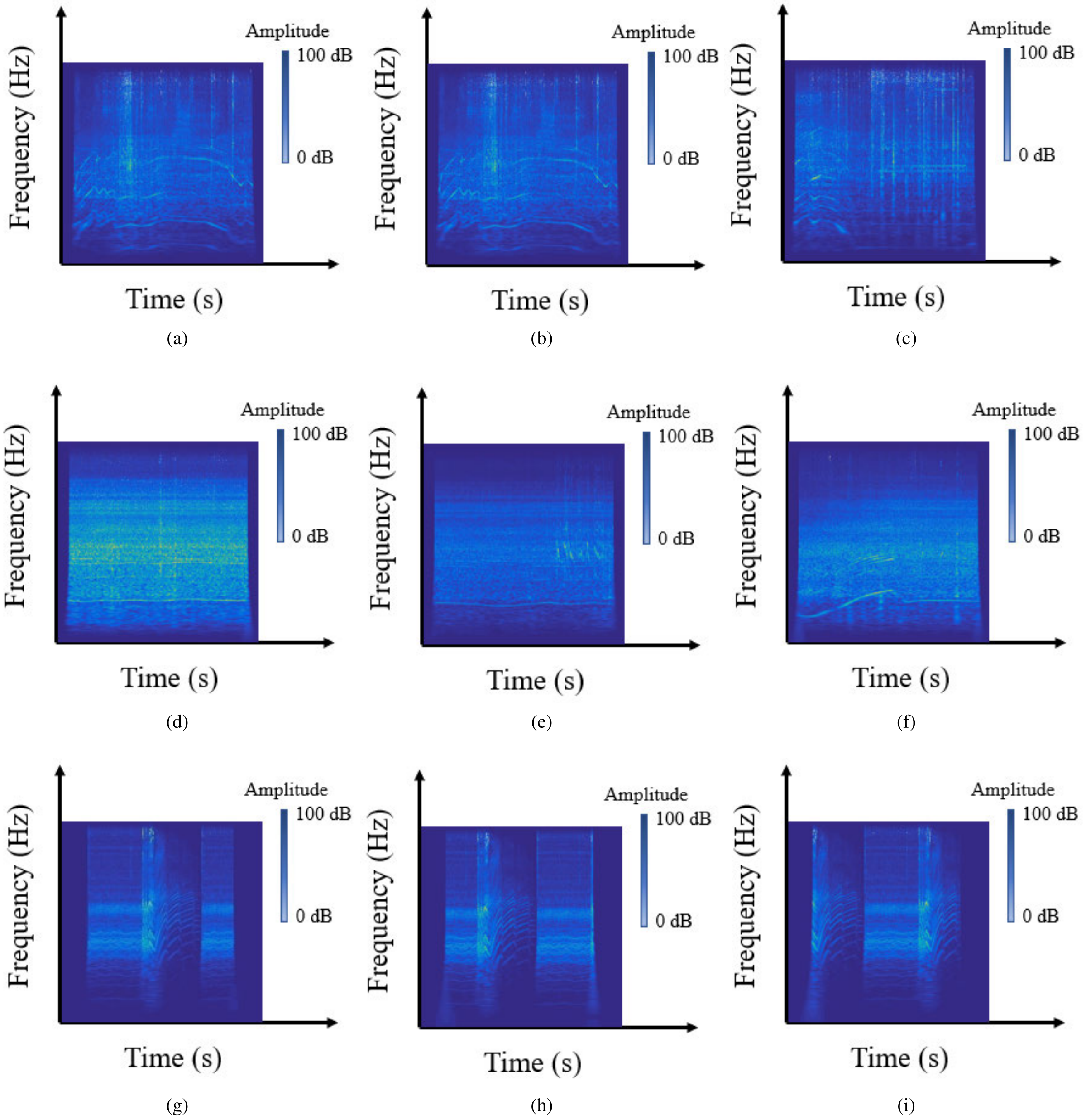


FIGURE 3. The illustration of three typical traffic acoustic scenes. Fig. 3a - Fig. 3c are the acoustic scenes of buses. Fig. 3d - Fig. 3f are the acoustic scenes of cars. Fig. 3g - Fig. 3i are the acoustic scenes of motorcycles. It can be found that these spectrograms are visually discriminated after constant Q transformation than the original audio clips.

and contrastive experimental screening, the best results are obtained when the number of gradient directions is 8 in 8 × 8 cells.

C. FEATURES EXTRACTED BY GFE AND LFE

The aforementioned LFE method describes the characteristics of HOG in a different point of view. Both of them improve the recognition accuracy of different acoustic scenes. The results are detailed in the Table 2. In Table 2, the method LFE represents the local feature extraction by Eq. 2. The method LFE only uses only the feature of 64 dimensions to achieve 73.98% accuracy. The LFE extended (LFEx) method concatenates the features extracted by Eq. 2 and Eq. 3. The GFE+LFE method concatenates the features extracted by the GFE and the LFE methods, whereas the GFE+LFEx method does with the features extracted by the GFE and the LFEx methods. The experimental results show the feature extraction method fused with GFE and LFEx achieves the best performance, and improves 4.23% of the accuracy comparing with the GFE method.

We also try different pooling methods, e.g. pooling over time domain and pooling over frequency domain. The experimental results are shown in Table 3. The numbers in the frequency column and time column labels depict the number of histograms for the frequency domain and time domain after pooling. For example, the first row presents all

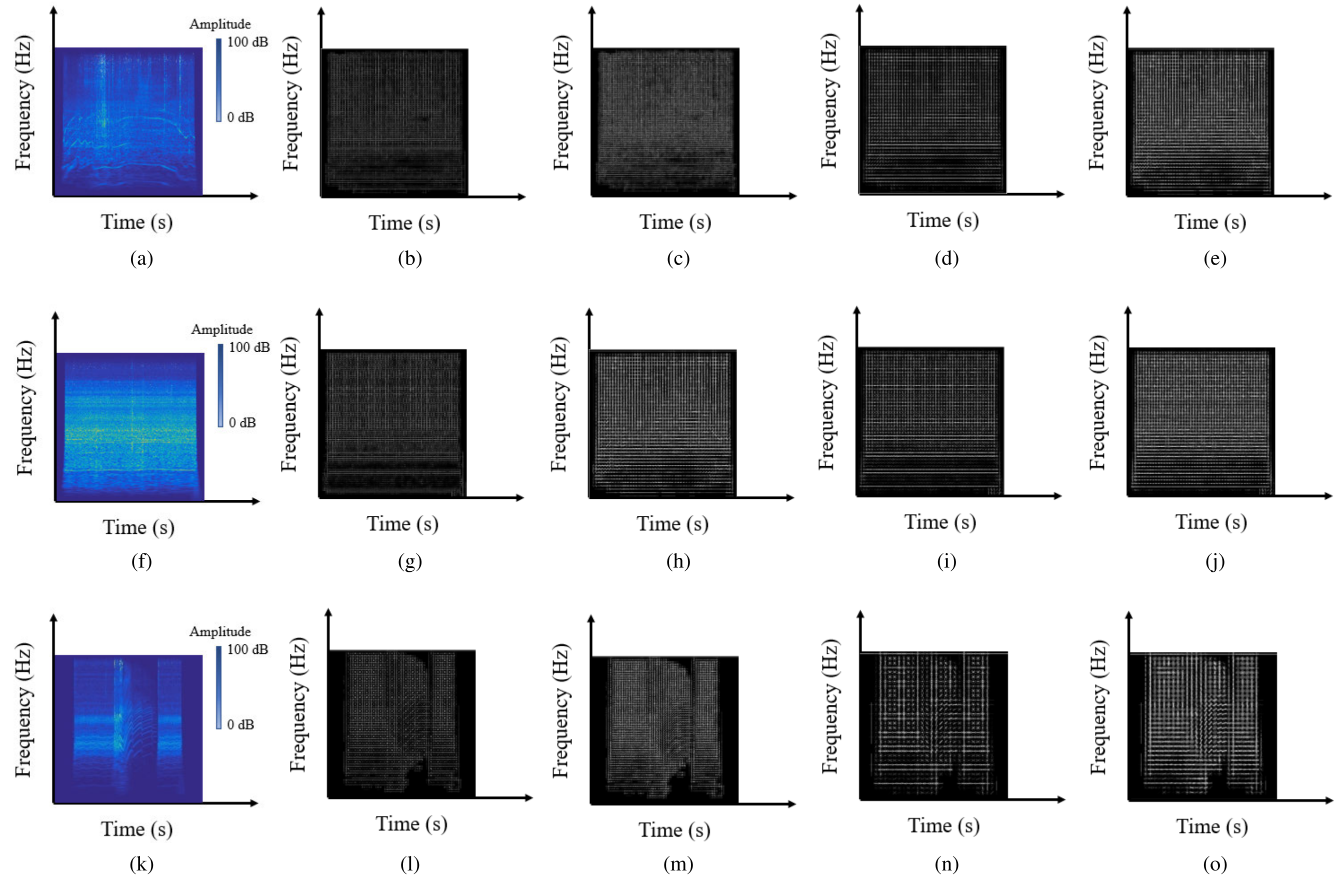


FIGURE 4. The illustration of CQT spectrograms and the corresponding HOG descriptors for three typical traffic acoustic scene. The first column presents the CQT spectrograms. The resolution of the cells in the second and third columns is 8 × 8, whereas the resolution of the fourth and fifth columns is 16 × 16. The number of gradient directions for the second and fourth columns is 8, whereas the number of gradient directions for the third and fifth columns is 32.

TABLE 2. The dimension of the extracted features for the GFE and LFE. Two simple but effective classifiers, e.g. linear support vector machine classifier (LinearSVC) and support vector machine classifier (SVC), are employed to evaluate the recognition accuracy of the GFE and LFE.

Methods Dimension	GFE 512	LFE 64	LFEx 128	GFE+LFE 512 + 64	GFE+LFEx 512 + 128
SVC (%)	84.73	72.96	78.81	87.13	88.96
LinearSVC (%)	86.53	73.98	81.69	86.64	86.66

TABLE 3. The recognition accuracy after pooling over time domain and frequency domain.

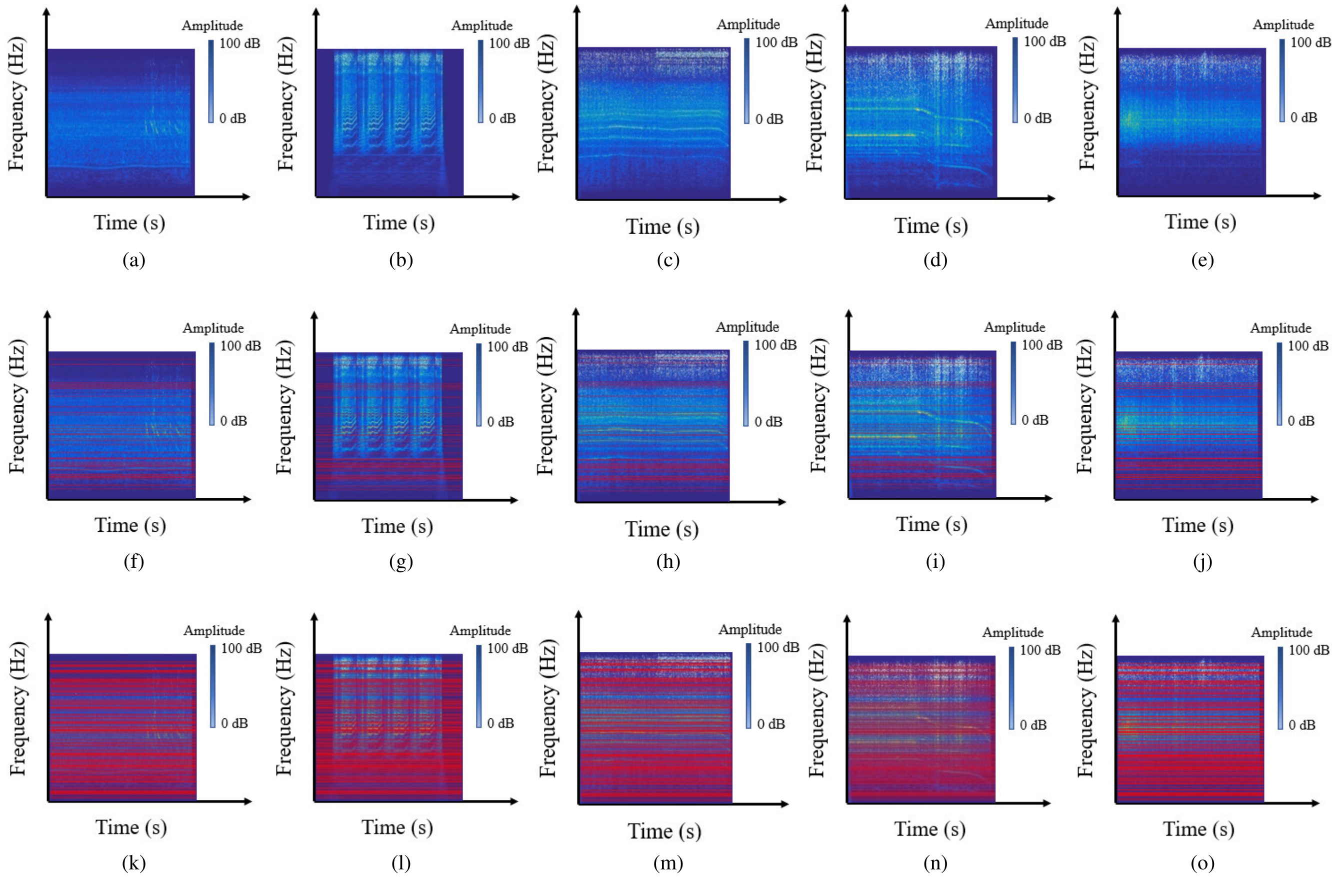
Frequency	Time	Dim	Accuracy (%)	
			SVC	LinearSVC
1	64	512	37.62	39.74
8	8	512	68.01	71.36
32	2	512	80.32	83.44
64	1	512	84.73	86.53

histograms have been averaged over the frequency domain. We find that pooling on the time domain achieves better performance.

D. PERFORMANCE EVALUATION

In this study, we employ LASSO to reduce the dimensions of feature vectors. In our implementation, the LASSO is employed from sklearn toolkit with default settings. Fig. 5 shows the selection of GFE features by LASSO. The results reveal that LASSO is an effective and efficient way to extract principle features from different acoustic scenes. Table 4 and Table 5 illustrate the comparison of the recognition accuracy by the GFE and LFE features, respectively. Table 4 lists the recognition accuracy by leveraging the features selected by LASSO on the GFE features, and Table 5 lists the recognition accuracy by leveraging the features selected by LASSO on the LFE features. In these two tables, the abbreviation *cv* means the number of holds for the cross-validation strategy. We choose different values of *cv* to demonstrate outperformance of our method under different of data distribution and the different accounts of data. The hyperparameter λ controls the shrinkage of the weighting parameters. In addition, the shrinkage parameter λ is also chosen by considering real-time requirement.

We also compare the recognition accuracy by leveraging the fused features extracted by the aforementioned method, e.g. GFE, LFE, and LASSO. The recognition accuracy



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

(k) (l) (m) (n) (o)

FIGURE 5. The illustration of the principle features selected by LASSO from five types of traffic acoustic scenes, e.g. car, motorcycle, bicycle, bus, and track. The red lines in the second and the third rows present the principle features selected. The first row the original CQT spectrograms. The second row presents 32 principle features, and the third one presents 231 principle features.

TABLE 4. The recognition accuracy by leveraging the features selected by LASSO on the GFE features. The abbreviation *cv* means the number of holds for the cross-validation strategy. The hyperparameter λ controls the shrinkage the weighting parameters.

Methods	GFE	LASSO ($cv = 5$)	LASSO ($cv = 20$)	LASSO ($cv = 10$)	GFE+LASSO ($cv = 10$)	GFE+LASSO ($\lambda = 2, cv = 10$)	GFE+LASSO ($\lambda = 10, cv = 10$)
Dimension	512	32	51	231	512+51	512	512
SVC (%)	84.73	77.39	78.81	81.63	84.63	86.63	85.12
LinearSVC (%)	86.53	74.68	77.63	83.34	86.46	87.01	85.53

TABLE 5. The recognition accuracy by leveraging the features selected by LASSO on the LFE features. The abbreviation *cv* means the number of holds for the cross-validation strategy. The hyperparameter λ controls the shrinkage the weighting parameters.

Methods	LFE	LASSO ($cv = 10$)	LASSO ($cv = 30$)	LASSO ($cv = 90$)	LFE+LASSO ($cv = 90$)	LFE+LASSO ($\lambda = 2, cv = 90$)	LFE+LASSO ($\lambda = 10, cv = 90$)
Dimension	128	32	34	79	128+79	128	128
SVC (%)	78.81	64.69	64.71	73.17	79.48	83.07	80.92
LinearSVC (%)	81.69	66.89	67.33	76.10	80.6	84.10	77.09

TABLE 6. The recognition accuracy by leveraging the features extracted by GFE, GFE+LFE, GFE+LASSO, and GFE+LFE+LASSO.

Methods	GFE+LFE	GFE+LASSO	GFE+LFE+LASSO
Dimension	512+128	512	512+128
SVC (%)	88.96	86.63	91.35
LinearSVC (%)	86.66	87.01	88.83

by leveraging the features extracted by GFE, GFE+LFE, GFE+LASSO, and GFE+LFE+LASSO is illustrated in Fig. 6. In Fig. 6, we find the GFE+LFE+LASSO almost achieves the best performance for all categories. The corresponding recognition accuracy is listed in Table 7. We find that the fused features achieve better results than the features by a single extraction method.

In addition, the local feature extraction methods are evaluated. The LASSO method makes a significant contribution to the final recognition accuracy. In Table 7, the comparisons of GFE, LFE, and GFE+LFE with different dimensions of features are given. The experimental results show that the LEF features contribute more than GFE features.

We compare the proposed method with two state-of-the-art methods. The first method is developed by Abidin *et al.* [29], which uses variable-Q transform (VQT) to generate the

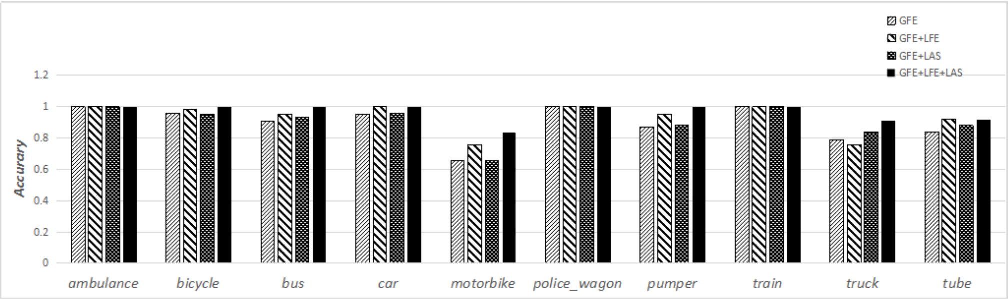


FIGURE 6. **The recognition accuracy by leveraging the features extracted by GFE, GFE+LFE, GFE+LASSO, and GFE+LFE+LASSO for each category of the traffic acoustic scenes in the proposed dataset.**

TABLE 7. **The comparison of the proposed method with state-of-the-art methods.**

Methods	Accuracy (%)
VQT+LBP [29]	85.50
Xception [30]	79.80
GFE+LFE	88.96
GFE+LASSO	87.01
GFE+LFE+LASSO	**91.35**

time-frequency representation for acoustic scene. Then, the adjacent evaluation completed local binary pattern (LBP) is adopted to extract time-frequency features. The second one is a deep learning method proposed by Yang *et al.* [30] for 2018 DCASE challenge. They extract multi-scale log-Mel features of acoustic signal. Then, a modified Xception network is developed to fuse the multi-scale features. These two methods are implemented and tested on our dataset. The first one achieves 85.5% of accuracy, whereas the second one does 79.8%. The comparison results are listed in Table 7. In Table 7, the three strategies of the proposed methods outperform the above the methods, because the GFE and LFE are effectively captures the acoustic features from the CQT spectrogram, and the LASSO can well select the discriminative features and eliminate the disturbance of the unrelated features.

E. DISCUSSION

The development of unmanned systems is inseparable from the effective awareness of the real-time, dynamic, and highly complex traffic environment. Traffic acoustic scene recognition (TASR) is a complex and challenging task aiming at recognizing acoustic environments solely based on an audio recording of the scene. Traffic acoustic scene recognition applications for unmanned vehicles can provide an auxiliary means besides of visual identification. The acoustic analysis and recognition, in consideration of its simple and convenient, can effectively enhance the perception ability which only applies visual information. These acoustic scenes can be defined according to specific geographical contexts, such as expressway, sidewalk, or metropolitan railway, and specific transportation tools, such as car, bus, or tramway. Accurate recognition of the scenes is really relevant for applications with the purpose of context machine awareness, which is of critical importance for intelligent transportation or automatic pilot. The feature extraction method presented is general, and can be extend to other time series analysis tasks, such traffic flow forecasting [31]–[33], intelligent computing [34]–[36], or medical signal visualization [37], [38].

IV. CONCLUSION

This article presents a new representation for traffic acoustic data, which accelerate and enhance traffic acoustic recognition. In order to achieve this, we transform the audio clip into the CQT spectrogram. The HOG descriptors are extracted from the CQT spectrogram. Then, a feature extraction method is proposed for both time domain and frequency domain on HOG descriptors. Two local feature extraction methods, which consider the volatility of the time-domain feature, are designed to describe the time-domain property. The dimension of the features is shrunk by LASSO to eliminate the negative affection inside the features in both the time domain and frequency domain. Furthermore, we collect sufficient real-world dataset for evaluating the performance of the proposed feature extraction method. The results on the real-world dataset demonstrate the outperformance of our two local feature extraction methods than the state-of-the-art frequency-domain feature extraction method.

The future work is conducting in two holds. First, we plan to use recurrent neural networks and its extension to extract more discriminated features, and use convolutional neural networks for more accurate recognition. Second, we would like to apply such a model to more complex and noisy roadway circumstances.

REFERENCES

[1] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," 2018, *arXiv:1807.09840*. [Online]. Available: http://arxiv.org/abs/1807.09840

[2] R. Serizel, V. Bisot, S. Essid, and G. Richard, "Machine listening techniques as a complement to video image analysis in forensics," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 948–952.

[3] V. Bisot, S. Essid, and G. Richard, "HOG and subband power distribution image features for acoustic scene classification," in *Proc. 23rd Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2015, pp. 719–723.

[4] M. Jessen, "Speaker classification in forensic phonetics and acoustics," in *Speaker Classification I*. Berlin, Germany: Springer, 2007, pp. 180–204.

[5] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 16–34, May 2015.

[6] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time–frequency representations for audio scene classification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 1, pp. 142–153, Jan. 2015.

[7] 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 Trans. Audio, Speech Language Process.*, vol. 14, no. 1, pp. 321–329, Jan. 2006.

[8] R. G. Malkin and A. Waibel, "Classifying user environment for mobile applications using linear autoencoding of ambient audio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2005, p. v-509.

[9] J. Moragues, A. Serrano, L. Vergara, and J. Gosálbez, "Acoustic detection and classification using temporal and frequency multiple energy detector features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 1940–1943.

[10] M. E. Niessen, T. L. Van Kasteren, and A. Merentitis, "Hierarchical sound event detection," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2013.

[11] M. Chum, A. Habshush, A. Rahman, and C. Sang, "IEEE AASP scene classification challenge using hidden Markov models and frame based classification," in *Proc. IEEE AASP Challenge Detection Classification Acoustic Scenes Events*, Oct. 2013.

[12] J. T. Geiger, B. Schuller, and G. Rigoll, "Large-scale audio feature extraction and SVM for acoustic scene classification," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2013, pp. 1–4.

[13] T. Andringa and J. Lanser, "How pleasant sounds promote and annoying sounds impede health: A cognitive approach," *Int. J. Environ. Res. Public Health*, vol. 10, no. 4, pp. 1439–1461, Apr. 2013.

[14] D. Jiang, K. Wu, D. Chen, G. Tu, T. Zhou, A. Garg, and L. Gao, "A probability and integrated learning based classification algorithm for high-level human emotion recognition problems," *Measurement*, vol. 150, Jan. 2020, Art. no. 107049.

[15] J. Lindström and C. Hanken, "Wearable computing: Security challenges, byod, privacy, and legal aspects," in *Wearable Technologies: Concepts, Methodologies, Tools, and Applications*. Hershey, PA, USA: IGI Global, 2018, pp. 1043–1067.

[16] G. Xiao, G. Tu, L. Zheng, T. Zhou, X. Li, S. H. Ahmed, and D. Jiang, "Multi-modality sentiment analysis in social Internet of Things based on hierarchical attentions and CSATTCN with MBM network," *IEEE Internet Things J.*, early access, Aug. 10, 2020, doi: 10.1109/JIOT.2020.3015381.

[17] D. Jiang, G. Tu, D. Jin, K. Wu, C. Liu, L. Zheng, and T. Zhou, "A hybrid intelligent model for acute hypotensive episode prediction with large-scale data," *Inf. Sci.*, vol. 546, pp. 787–802, Feb. 2021.

[18] K. Patil and M. Elhilali, "Multiresolution auditory representations for scene classification," *Cortex*, vol. 87, no. 1, pp. 516–527, 2002.

[19] K. Lee, Z. Hyung, and J. Nam, "Acoustic scene classification using sparse feature learning and event-based pooling," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2013, pp. 1–4.

[20] H. Lu, D. Huang, Y. Song, D. Jiang, T. Zhou, and J. Qin, "ST-TrafficNet: A spatial-temporal deep learning network for traffic forecasting," *Electronics*, vol. 9, no. 9, p. 1474, Sep. 2020.

[21] J. Ye, T. Kobayashi, M. Murakawa, and T. Higuchi, "Acoustic scene classification based on sound textures and events," in *Proc. 23rd ACM Int. Conf. Multimedia (MM)*, 2015, pp. 1291–1294.

[22] A. Dessein, A. Cont, and G. Lemaitre, "Real-time detection of overlapping sound events with non-negative matrix factorization," in *Matrix Information Geometry*. Berlin, Germany: Springer, 2013, pp. 341–371.

[23] H. Phan, L. Hertel, M. Maass, P. Koch, and A. Mertins, "Label tree embeddings for acoustic scene classification," in *Proc. 24th ACM Multimedia Conf. (MM)*, 2016, pp. 486–490.

[24] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc., B, Methodol.*, vol. 58, no. 1, pp. 267–288, Jan. 1996.

[25] J.-M. Azaïs, Y. De Castro, and S. Mourareau, "Power of the spacing test for least-angle regression," *Bernoulli*, vol. 24, no. 1, pp. 465–492, Feb. 2018.

[26] J. C. Brown, "Calculation of a constant q spectral transform," *J. Acoust. Soc. Amer.*, vol. 89, no. 1, pp. 425–434, Jan. 1991.

[27] J. C. Brown and M. S. Puckette, "An efficient algorithm for the calculation of a constant q transform," *J. Acoust. Soc. Amer.*, vol. 92, no. 5, pp. 2698–2701, Nov. 1992.

[28] X. Li, L. Bai, Z. Ge, Z. Lin, X. Yang, and T. Zhou, "Early diagnosis of neuropsychiatric systemic lupus erythematosus by deep learning enhanced magnetic resonance spectroscopy," *J. Med. Imag. Health Informat.*, vol. 11, no. 2, 2021.

[29] S. Abidin, R. Togneri, and F. Sohel, "Spectrotemporal analysis using local binary pattern variants for acoustic scene classification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 2112–2121, Nov. 2018.

[30] L. Yang, X. Chen, and L. Tao, "Acoustic scene classification using multiscale features," in *Proc. Detection Classification Acoustic Scenes Events (DCASE)*, 2018, pp. 29–33.

[31] W. Cai, J. Yang, Y. Yu, Y. Song, T. Zhou, and J. Qin, "PSO-ELM: A hybrid learning model for short-term traffic flow forecasting," *IEEE Access*, vol. 8, pp. 6505–6514, 2020.

[32] L. Cai, Y. Yu, S. Zhang, Y. Song, Z. Xiong, and T. Zhou, "A sample-rebalanced outlier-rejected k-nearest neighbor regression model for short-term traffic flow forecasting," *IEEE Access*, vol. 8, pp. 22686–22696, 2020.

[33] L. Cai, M. Lei, S. Zhang, Y. Yu, T. Zhou, and J. Qin, "A noise-immune LSTM network for short-term traffic flow forecasting," *Chaos*, vol. 30, no. 3, pp. 1–10, 2020.

[34] D. Jiang, Z. Liu, L. Zheng, and J. Chen, "Factorization meets neural networks: A scalable and efficient recommender for solving the new user problem," *IEEE Access*, vol. 8, pp. 18350–18361, 2020.

[35] J. Wang, Z. Xie, Y. Li, Y. Song, J. Yan, W. Bai, T. Zhou, and J. Qin, "Relationship between health status and physical fitness of college students from south China: An empirical study by data mining approach," *IEEE Access*, vol. 8, pp. 67466–67473, 2020.

[36] L. Zheng, N. Guo, J. Yu, and D. Jiang, "Memory reorganization: A symmetric memory network for reorganizing neighbors and topics to complete rating prediction," *IEEE Access*, vol. 8, pp. 81876–81886, 2020.

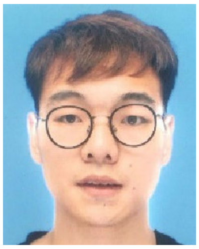
[37] T. Zhou, G. Han, B. N. Li, Z. Lin, E. J. Ciaccio, P. H. Green, and J. Qin, "Quantitative analysis of patients with celiac disease by video capsule endoscopy: A deep learning method," *Comput. Biol. Med.*, vol. 85, pp. 1–6, Jun. 2017.

[38] B. N. N. Li, X. Wang, R. Wang, T. Zhou, R. Gao, E. J. Ciaccio, and P. H. Green, "Celiac disease detection from videocapsule endoscopy images using strip principal component analysis," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Nov. 15, 2019, doi: 10.1109/TCBB.2019.2953701.

DAZHI JIANG received the B.A. degree in computer science from the China University of Geosciences, Wuhan, in 2004, and the Ph.D. degree from the State Key Laboratory of Software Engineering, Wuhan University, China, in 2009. Since 2009, he has been with the Department of Computer Science, Shantou University, China, where he was a Professor. His research interests include affective computing, deep learning, data mining, and applications of artificial intelligence.

DONGMIN HUANG is currently a Graduate Student with the Department of Computer Science, Shantou University. His current research interests include EEG-base affective computing, signal processing, and deep learning.

YOUYI SONG is currently pursuing the Ph.D. degree with the Centre of Smart Health, The Hong Kong Polytechnic University. His research interests include clinical science, medical image segmentation, machine learning, and data analysis.



KAICHAO WU is currently pursuing the master's degree with the Department of Computer Science, Shantou University, China. His research interests include image processing, machine learning, and time series analysis.

HUAKANG LU is currently pursuing the bachelor's degree with the Department of Computer Science, Shantou University, China. His research interests include intelligent transportation systems, machine learning, and computer vision.

QUANQUAN LIU received the master's degree from the Department of Computer Science, Shantou University, China. His research interests include machine learning and time series analysis.

TENG ZHOU is currently an Assistant Professor with the Department of Computer Science, Shantou University, and also serves as a Research Associate with the Center of Smart Health, The Hong Kong Polytechnic University. His research interests include intelligent transportation systems and machine learning.