


Article

Acoustic Scene Classification Using Efficient Summary Statistics and Multiple Spectro-Temporal Descriptor Fusion

Jiaxing Ye ^{1,*} , Takumi Kobayashi ², Nobuyuki Toyama ¹, Hiroshi Tsuda ¹
and Masahiro Murakawa ²

¹ National Metrology Institute of Japan (NMIJ), National Institute of Advanced Industrial Science and Technology (AIST), 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568 Japan; toyama-n@aist.go.jp (N.T.); hiroshi-tsuda@aist.go.jp (H.T.)

² Artificial Intelligence Research Center (AIRC), National Institute of Advanced Industrial Science and Technology (AIST), 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568 Japan; takumi.kobayashi@aist.go.jp (T.K.); m.murakawa@aist.go.jp (M.M.)

* Correspondence: jiaxing.you@aist.go.jp; Tel.: +81-29-861-2850

Received: 12 February 2018; Accepted: 9 August 2018; Published: 13 August 2018

Abstract: This paper presents a novel approach for acoustic scene classification based on efficient acoustic feature extraction using spectro-temporal descriptors fusion. Grounded on the finding in neuroscience—“auditory system summarizes the temporal details of sounds using time-averaged statistics to understand acoustic scenes”, we devise an efficient computational framework for sound scene classification by using multiple time-frequency descriptors fusion with discriminant information enhancement. To characterize rich information of sound, i.e., local structures on the time-frequency plane, we adopt 2-dimensional local descriptors. A more critical issue raised in how to logically ‘summarize’ those local details into a compact feature vector for scene classification. Although ‘time-averaged statistics’ is suggested by the psychological investigation, directly computing time average of local acoustic features is not a logical way, since arithmetic mean is vulnerable to extreme values which are anticipated to be generated by interference sounds which are irrelevant to the scene category. To tackle this problem, we develop time-frame weighting approach to enhance sound textures as well as to suppress scene-irrelevant events. Subsequently, robust acoustic feature for scene classification can be efficiently characterized. The proposed method had been validated by using Rouen dataset which consists of 19 acoustic scene categories with 3029 real samples. Extensive results demonstrated the effectiveness of the proposed scheme.

Keywords: acoustic scene classification; time-frequency analysis; local descriptor; summary statistics; convex combination

1. Introduction

Environmental sounds, which are an integral part of multimedia data, contain plenty of information, such as location and activities. To efficiently utilize the audio information for indexing massive multimedia contents, many research efforts have been spent on developing acoustic scene classification (ASC) system using advanced signal processing and machine learning techniques in recent years [1–3]. Although some progress has been made, the key issues in acoustic scene understanding, i.e., acoustic feature representations development and efficient framework, are still open questions to the research field.

Psychological research findings reveal that “auditory system summarizes the temporal details of sounds using time-averaged statistics to understand acoustic scenes” [4,5]. Inspired by the results,

plenty of works endeavour to adopt descriptive statistics to characterize textures in an acoustic scene for content-based classification [2,6]. Standard approaches to ASC firstly convert input audio signal to a time-frequency representations (TFRs) by using either handcrafted features, such as Mel-scale spectrogram and mel-frequency cepstral coefficients (MFCCs), or learning-based feature, e.g., unsupervised feature learning with neural networks. Then, statistical moments, e.g., mean, variance, skewness and kurtosis, are subsequently employed to convert TFRs (matrix) to compact feature vector [7–9] for further statistical classification using supervised learning. Figure 1 shows a general flowchart of ASC.

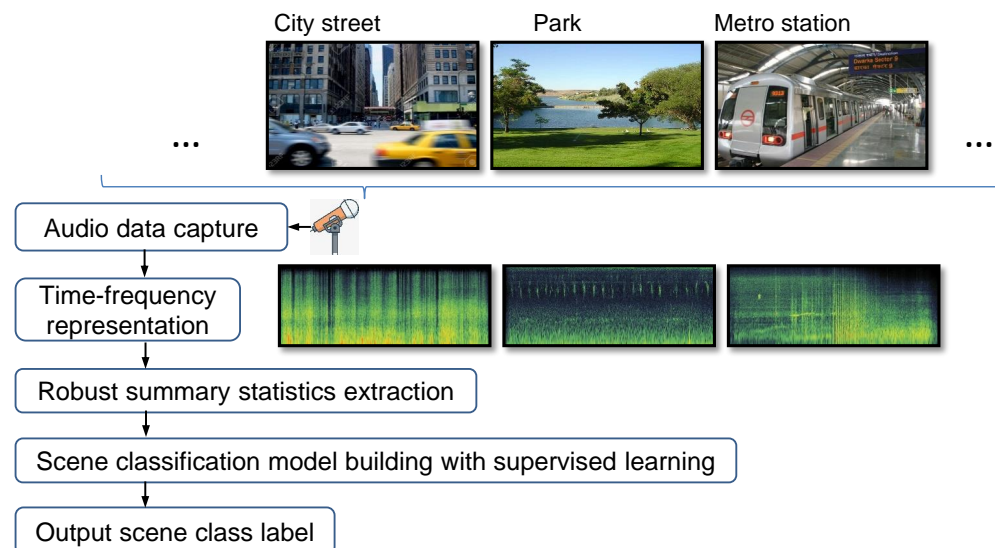


Figure 1. General framework of acoustic scene classification system. ASC can be performed by examining homogeneous structures of audio data.

Although descriptive statistics are off-the-shelf tools which can be used for acoustic scene understanding, the “summarization” performed by the auditory system is essentially different from arithmetic averaging. Descriptive statistics of time-averaging implies that every frame of acoustic feature carries similar information for ASC. In contrast, the auditory system adopts an adaptive scheme that “keeping the environment-specific acoustic information, while weeding out the irrelevant detail”, according to the neuroscience research result [10]. In other words, auditory system has a *content-enhancing* function that draws more attention on discriminative patterns in acoustic scenes, which is well-suited to acoustic scenes parsing.

This paper attempts to mimic *discriminantcontent – enhancing function* of auditory perception for ASC by using data-driven statistical machine learning. We begin this study with the primary mathematical formulation, that is, we assume ambient sound can be decomposed into two categories, which can be expressed as follows:

$$s(t) = s_r(t) + s_i(t). \quad (1)$$

Concretely in ASC tasks, $s_r(t)$ and $s_i(t)$ denote scene-relevant/irrelevant sounds, respectively. The first-category usually exhibits high temporal homogeneity, and thus delivers predominant discriminative information of acoustic scenes. Furthermore, it can be described as a superposition of many similar scene-relevant acoustic events over background textures [4], which have been extensively investigated in psychology research of human auditory perception [7,11]. The latter component $s_i(t)$ denotes scene-irrelevant sound which hardly contributes to ASC. For instance, speech can occur in different acoustic scenes, such as in street, shop, and cafe. It is noteworthy that $s_i(t)$ usually presents complex spectro-temporal patterns and stronger energy, and thus can severely affect ASC performance. Therefore, these outliers to current acoustic scene category should be handled carefully at feature

extraction stage. By exploiting characteristics of the two components, we found that $s_r(t)$ commonly present stationary statistical properties over time; in contrast, scene-irrelevant sounds, which exhibit complex spectro-temporal patterns in short-time periods, are discretely superimposed. In this study, we investigate the structural difference to discern s_i from s_r . Main contributions of this work are listed as follows.

- Latest research towards ASC manifested that 2-dimensional (2D) local descriptors are efficient for describing environmental sounds, such as using local Binary patterns (LBP) [12] and histograms of oriented gradients (HOG) [13]. We perform intensive tests to evaluate various local descriptors for ASC. Furthermore, we proposed a framework to aggregate multiple 2D descriptors for ASC.
- To enhance scene-specific sound patterns, we conduct novelty detection over the audio clip. Both sound textures and super-positioned scene-relevant events would reside in the subspace due to high temporal homogeneity. On the contrary, scene-irrelevant sounds will generate distinct deviations to the subspace. According to above analysis, a series of weights can be derived which indicate the importance of representing the scene.
- To efficiently summarize local acoustic patterns, we employ a weighted averaging scheme which converts spectro-temporal distribution (matrix) to a compact vector. A multi-feature aggregation scheme had been further applied to fuse the discriminant information conveyed by local descriptors. According to the validation studies on real data, the proposed approach achieved superior performance comparing to other recent results.

2. Related Works

In computational auditory scene analysis, standard approaches are primarily composed of two steps: feature extraction and statistical classification. The first stage tackles the issue of developing efficient feature representation of an audio clip for ASC. A wide variety of acoustic features had been investigated, including log-scale spectrogram [1], cepstral features (MFCCs) [6] and gammatone cepstral coefficients (GTCC) [2]. Those features are mainly taken from automatic speech recognition (ASR) field and have been proved to be effective to characterize rich low-level information from the audio signal by using a time-frequency (matrix) representation (TFR). Subsequently, statistical moments, e.g., averaging and standard deviations, are employed to convert time-frequency representations (matrix) to compact feature vector [2]. Besides, various types of summary statistics had been adopted to characterize discriminant information in both time and frequency domain, such as by using zero-crossing rate (ZCR) and spectral kurtosis [3]. In recent years, much attention is paid to apply advanced feature extraction methods for ASC, such as using Bag-of-features [13] and sparse coding models [1] to distill discriminant information from noisy recordings. At the latter part, the typical examples of statistical classification algorithms are Gaussian mixture models (GMM) and support vector machine (SVM) [14]. More recently, taking inspiration from various successful applications in both computer vision and speech recognition, there is an emerging trend in ASC research to shift from conventional classification techniques to deep neural network-based methods. Such tendency is evident according to the latest DCASE challenge series (IEEE AASP Challenges Detection and Classification of Acoustic Scenes and Events) [15]. Many researchers attempted to introduce Convolutional Neural Networks (CNNs) to drive an informative and robust audio data representation for ASC in a data-driven manner, such as in [16,17]. Specifically, the CNNs-based approach is able to jointly optimize the acoustic feature representation and the classification algorithm. However, the current open datasets for ASC research are much smaller compared with the ones for computer vision, which may lead to under fitting status. Besides, the ensemble of several ASC architectures had been proved to be efficient to boost the ASC accuracy. For instance, in 2017 DCASE evaluation, the well-ranked submissions were based on CNNs aggregating with other deep neural network models, i.e., Recurrent Neural Network (RNN) and Multiple Layer Perception (MLP) [18]. Current work had been carried out based on the survey over state-of-the-art research.

In this study, we propose novel ASC approach which consists of three major components: spectro-temporal feature extraction using multiple time-frequency representations (TFRs), discriminant content enhancing weights extraction and multi-descriptor aggregation scheme for ASC. Figure 2 shows the processing flow of the proposed ASC system. We introduce the key components as follows.

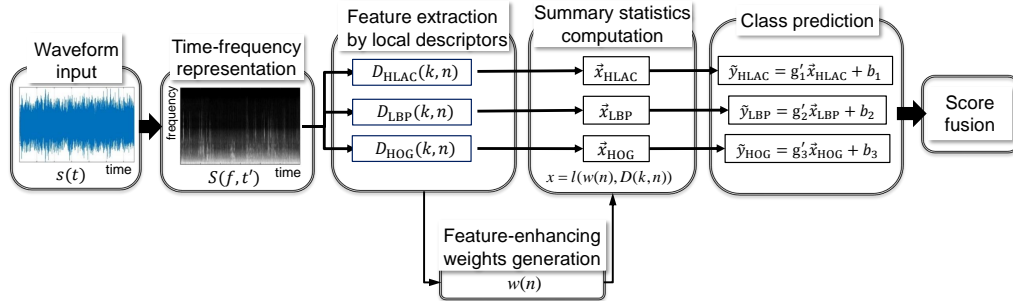


Figure 2. Flowchart of the proposed acoustic scene classification approach.

3. Proposed Method

3.1. Time-Frequency Representation (TFR)

Audio waveform is commonly transformed to time-frequency representations (TFRs) in which temporal and spectral information can be characterized simultaneously. In this study, we evaluated several TFRs, including Mel-spectrogram, MFCCs [2] and Constant-Q spectrogram [19]. We denote TFRs as $S(f, t')$ in Figure 2, where t' is frame index.

3.2. Spectro-Temporal Descriptors

Based on TFRs, we further employ local descriptors to characterize spectro-temporal patterns in a 2D fashion. Several local descriptors are evaluated, such as Higher-order Local Auto-Correlation (HLAC), Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG). By using descriptors, we convert raw TFRs to mid-level local structure-based representations which facilitate ASC. In Figure 2, the feature extracted by using local descriptors is noted by $D(n, k)$, where n is time index and k is feature dimension.

3.2.1. Higher-order Local Auto-Correlation (HLAC)

Higher-order Local Auto-Correlation (HLAC) features are conventional local descriptors for extracting patterns in 2D patch [20]. The features had been successfully applied to a wide variety of real applications, including texture and face classification. The HLAC features is well-developed based on higher-order autocorrelation function:

$$D_{\text{HLAC}}(\mathbf{a}_1, \mathbf{a}_2) = \int S(\mathbf{r})S(\mathbf{r} + \mathbf{a}_1)S(\mathbf{r} + \mathbf{a}_2) \quad (2)$$

The mask patterns of HLAC is shown in Figure 3. In dealing with audio, $S(r)$ denotes TFRs, $\mathbf{r} = [t_r, f_r]^\top$ is reference point on time-frequency plane, $(\mathbf{a}_1 = [t_{a1}, f'_{a1}]^\top, \mathbf{a}_2 = [t_{a2}, f'_{a2}]^\top)$ is a set of displacements. HLAC extraction is limited to 3×3 local region and there are 35 individual mask patterns extracted. We introduce a sliding window on TFRs covering 3 consecutive frames of spectrum, from where HLAC features are extracted. The window shifts one frame at a time. Since acoustic features are assumed to be highly correlated within local region, more discriminative features can to be obtained via computing HLAC.

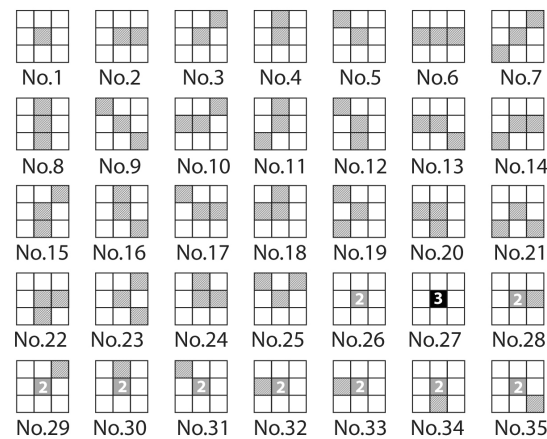


Figure 3. HLAC coding scheme on spectro-temporal local regions.

3.2.2. Local Binary Patterns (LBP)

Local binary patterns (LBP) are effective local descriptors which have been applied for textures classification [21] and sound classification [12], etc. The LBP convert local structures into binary patterns by comparing values to the central pixel, which is briefly introduced by Figure 4. In this study, we adopt LBP as spectro-temporal feature extractor. The general formulation for LBP can be written as:

$$D_{\text{LBP}}(\mathcal{L}_{\mathbf{c}}; \tau_{\mathbf{c}}) = \sum_{j=1}^J 2^{j-1} \llbracket I(\mathbf{r}_j) > \tau_{\mathbf{c}} \rrbracket \quad (3)$$

where I is gray-level pixel value at spatial position \mathbf{r}_j , and $\llbracket \cdot \rrbracket$ generates 1 only if bracketed condition is met and 0 otherwise. $\mathcal{L}_{\mathbf{c}} = \{\mathbf{r}_j\}_{j=1}^J$ indicates local 2D structure surrounded $\mathbf{c} \in \mathcal{R}^2$, including J spatial positions \mathbf{r}_j close to \mathbf{c} and $\tau_{\mathbf{c}}$ is the gray-level value of the central pixel. For ordinary LBP, J is set to 8 and hence local patch is limited to 3×3 . In this study, we use prototype version.

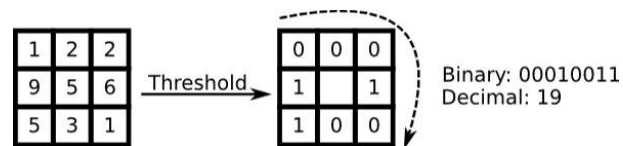


Figure 4. LBP coding scheme on spectro-temporal local regions.

3.2.3. Histogram of Oriented Gradients (HOG)

Histogram of oriented Gradients (HOG) is one most important 2D local descriptor in image processing, which count occurrences of gradient orientation in the localized patch. It has also been successfully applied for sound processing [13]. In a similar vein, we introduce HOG descriptor to characterize spectro-temporal structures in acoustic scenes with cumulative oriented gradients over local TFR. Figure 5 shows primary mechanism of HOG features, which extract plenty of spectro-temporal dynamics on time-frequency plane. The extracted HOG acoustic features are favorable for characterizing wide variations in environmental sound.

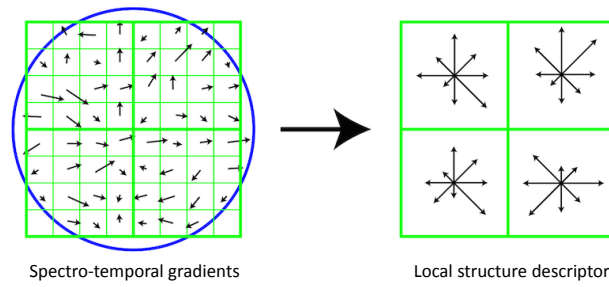


Figure 5. HOG coding scheme on spectro-temporal local regions.

3.3. Acoustic Summary Statistics Extraction for ASC

Although spectro-temporal descriptors can extract rich details of acoustic signal, more critical issue can be raised in summarizing those local patterns to derive a compact feature vector for classification. In this study, we devise an efficient approach to extract robust summary statistics from audio scene data. First, we present an algorithmic flowchart in Algorithm 1, and the details are demonstrated as follows.

Algorithm 1 Texture-enhancing weights generation algorithm

```

1: procedure SUMMARYSTATEXT( $\mathbf{D}, \varepsilon_{kurt}, \varepsilon_{skew}$ )
2:   Perform eigen decomposition to acoustic feature  $\mathbf{D}$  using (5)
3:   Compute novelty degree  $\mathbf{h}$  using (6)
4:   Extract Gaussian measures  $v_{skew}, v_{kurt}$ 
5:   if  $v_{kurt} > \varepsilon_{kurt} || v_{skew} > \varepsilon_{skew}$  then ▷ Event detected
6:     Likelihood extraction with Laplacian model using (7)
7:      $\mathbf{w} \leftarrow \text{Lap}(\mathbf{h} | \mu_{\mathbf{h}}, b_{\mathbf{h}})$  ▷ Weights generation
8:   else ▷ No event detected
9:      $\mathbf{w} \leftarrow \mathbf{1}_N$  ▷ Uniform weights applied
10:  end if
11:  Summary statistics extraction with weighted averaging (9)
12:  return  $\mathbf{x}$ 
13: end procedure

```

3.3.1. Unsupervised Novelty Analysis of Acoustic Scene

Our goal is to enhance scene relevant acoustic patterns and to suppress scene-irrelevant events as well during summary statistics extraction. To this end, it is necessary to discern the two components in environmental sound clips. According to (1), two category sounds are linearly combined, we adopt (linear) subspace method to detect events in an acoustic scene in an unsupervised manner. Sound textures, which are composed of the superposition of many acoustic events with high similarity, will reside in principal acoustic subspace; on the contrary, scene-irrelevant events are anticipated to exhibit distinct distance to the subspace. The procedure is based on principal component analysis (PCA) [22] and we start the process from computing correlation matrix of input feature:

$$\mathbf{C}_{\mathbf{D}} = \frac{1}{N} \sum_{n=1}^N \mathbf{d}_n \mathbf{d}_n^{\top} \quad (4)$$

where $\mathbf{d}_n, n \in [1, N]$ is acoustic feature vector extracted from one audio sample, i.e. 30 s clip, then, eigen decomposition is performed:

$$\lambda \mathbf{v} = \mathbf{C}_{\mathbf{D}} \mathbf{v} \quad (5)$$

Let $\mathbf{P} = [\mathbf{v}_1, \dots, \mathbf{v}_{K'}]$ denotes subspace accommodating predominant textures, which is composed of K' -th eigen vectors with highest eigenvalues. K' is determined by contribution rate which is defined as $\eta'_K = \sum_{k=1}^{K'} \lambda_k / \sum_{k=1}^K \lambda_k$. The deviation distance to subspace can be computed by:

$$h = \mathbf{d}^\top \mathbf{d} - \mathbf{d}^\top \mathbf{P} \mathbf{P}^\top \mathbf{d} \quad (6)$$

By examining residual h , we are able to detect outlier events in the acoustic scene. Sound textures, due to high temporal homogeneity, will generate h obeying normal distribution. In contrast, scene-irrelevant events superimposed on textures will introduce long tail to the histogram of h , therefore, h can no longer be well described by Gaussian. Based on such property, we introduce Gaussianity measures of kurtosis and skewness to discern acoustic events, which are denoted as $v_{kurt}(h)$ and $v_{skew}(h)$, respectively. Thresholds on two measures, which are $\varepsilon_{kurt}, \varepsilon_{skew}$, are experimentally set to detect scene-irrelevant events. If there are no events detected by thresholding, uniform weights can be applied since there are mostly homogeneous textures. Otherwise, we develop weights to suppress scene-irrelevant events as follows.

3.3.2. Textures-Enhancing Weights Generation

To enhance sound textures for ASC, we develop weights for feature frames based on membership probabilities to the acoustic scene. Laplace model, due to its robustness to outliers, is introduced to derive such probabilities [22]. By fitting h to the model, μ_h, b_h and likelihood can be estimated:

$$\text{Lap}(h|\mu_h, b_h) = \frac{1}{2b_h} \exp\left(-\frac{|h - \mu_h|}{b_h}\right) \quad (7)$$

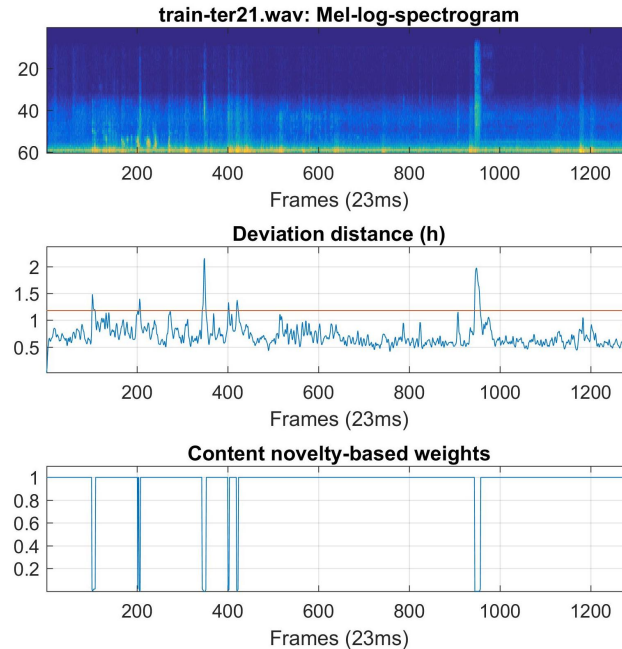


Figure 6. Example of acoustic texture enhancing weights generation.

Laplace model puts much less probability density to events which are unrelated to scene category. To further enhance sound textures, we introduce soft thresholding:

$$w = \begin{cases} \text{Lap}(h|\mu_h, b_h) & h > \tau \\ 1 & h \leq \tau \end{cases} \quad (8)$$

where $\tau = \mu_h + \sqrt{2b_h}$. Such setting assures textures will obtain highest weights (1, from probability aspect) while scene-irrelevant events would be suppressed by smaller weights. In Figure 6, we present an example of textures-enhancing weight generation using proposed method.

3.3.3. Summary Statistics Computation

Based on the weights $w(n)$ and feature representation $D(n, k)$, we derive summary statistics by:

$$\mathbf{x} = l(\mathbf{D}, \mathbf{w}) = \sum_{n=1}^N w(n) \mathbf{d}_n \quad (9)$$

where \mathbf{d}_n is the n -th feature vector in $\mathbf{D} \in \mathcal{R}^{N \times K}$ and \mathbf{x} is the extracted feature vector.

3.4. Class Score Fusion for Classification

As shown in Figure 2, we fuse discriminant information characterized by multiple spectro-temporal descriptors for ASC. To this end, we estimate class membership probabilities of input sound clip using various acoustic features with probabilistic SVM, which generates the probability interpretation of distance between input data and classification hyperplane in the (kernel) feature space. The formulation can be expressed as:

$$\min_{A, B} \frac{1}{M} \sum_{m=1}^M \log (1 + \exp (-y_m (A(\mathbf{w}'_{svm} \Phi(\mathbf{x}_{HLAC, m} + b_{svm}) + B))) \quad (10)$$

where $\{\mathbf{x}_{HLAC, m}, y_m\}$ are the HLAC feature vector extracted from m -th training clip and the corresponding label, respectively. Parameters of \mathbf{w}_{svm} and b_{svm} can be determined by quadratic programming, and logistic regression can be performed to compute A, B accordingly. Finally, we can derive class score $l_{HLAC, m}$. In the same vein, the class probabilities of $l_{LBP, m}$ and $l_{HOG, m}$ can be computed by using features $\{\mathbf{x}_{LBP, m}, y_m\}$ and $\{\mathbf{x}_{HOG, m}, y_m\}$. Finally, we employed linear programming to calibrate multi-stream class scores as follows:

$$l_{SCENE} = (\alpha_1) \times l_{HLAC} + (\alpha_2) \times l_{LBP} + (\alpha_3) \times l_{HOG}, \quad 0 \leq \alpha_i \leq 1, i \in [1, 2, 3], \quad \sum_{i=1}^3 \alpha_i = 1 \quad (11)$$

in which the conditional fusion weights α_i can be tuned explicitly at the training/validation stage. The estimated score fusion formula is anticipated to achieve higher accuracy comparing to the case of simple majority voting.

4. Experiments

4.1. Dataset and Parameters

We validate the proposed scheme by performing extensive experiments using LITIS Rouen Dataset [14], which includes 19 classes of real acoustic scenes categories with 3026 clips of 30 s length. It is noteworthy that the LITIS Rouen dataset is open dataset, which provide a standardised way to present and compare results. Figure 7 shows the distribution of audio samples among classes. Sounds are recorded at 22.05kHz sampling rate and 16-bit depth. 20-fold splits are provided to partition data into 80%-training/20%-test sets. In our experiments, we set Fourier analysis window length to 30ms with half overlapping. 60 Mel-filters were applied to extract Mel-spectrogram. To obtain CQT spectrogram, the number of bins per octave was set to 48 and local region size, number of orientations were set to 8×8 and 8, respectively. 2D smoothing was performed to TFRs through convolving with Gaussian kernel and kernel parameter was set to 3. At acoustic subspace extraction, we set contribution rate η'_K to 0.99. Thresholds for events detection of ε_{kurt} and ε_{skew} were set to 7 and 1, respectively. At classification stage, Gaussian kernel parameter was set to 0.1.

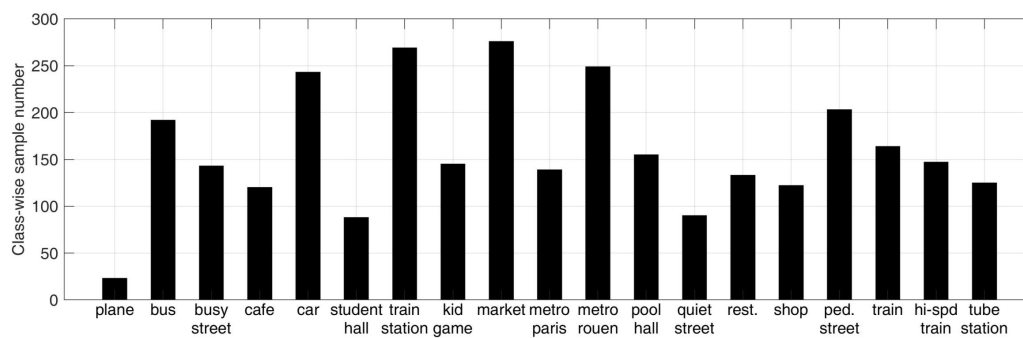


Figure 7. Statistics of acoustic scene recordings collection by categories in Rouen dataset.

4.2. Evaluation of TFRs with Local Descriptors

We began experiments by testing acoustic features extracted by local descriptors over various time-frequency representations (TFR). In detail, we examined four kinds of TFRs, including Fourier spectrogram, Mel-scaled spectrogram (Mel-spectrogram), constant-Q transform spectrogram (CQT) and Mel-Frequency Cepstral Coefficients (MFCCs), due to their popularity in ASC research. Furthermore, well-developed two-dimensional local descriptors of HLAC, LBP and HOG had been introduced to further characterize spectro-temporal patterns on time-frequency plane. Together, we had 12 combinations of TFRs with local descriptors. Since this evaluation was dedicated to acoustic feature comparison, simple time-averaging was applied to produce feature vector for each audio clip, i.e., \mathbf{w} was set to $\mathbf{1}_N$ in (9); SVM classifier with Gaussian kernel had been used for multi-class classification. Table 1 summarized all the results, from which we can see the highest ASC precision was achieved by using HOG descriptor over the CQT spectrogram. The performance comparison also revealed that although Mel-spectrogram and MFCCs were widely applied, they are not optimal when working with local descriptors because much local detail information was lost during feature extraction. In addition, we present the class-wise feature distribution of the LITIS Rouen dataset by using CQT spectrogram with HoG features in Figure 8. According to the feature space visualization, the compact clusters are anticipated to achieve higher ASC accuracy, while scattered clusters can be difficult to classify.

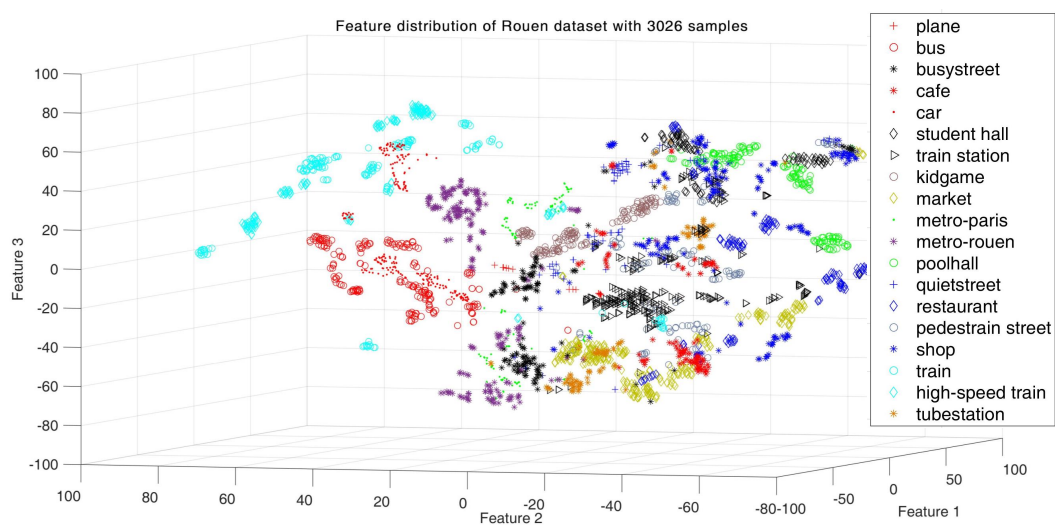


Figure 8. Visualization of Rouen dataset by using CQT+HoG features and t-SNE [23].

Table 1. ASC Results by using TFRs with local descriptors.

	Spectro.	Mel-Spectro.	CQT	MFCC
raw	77.38%	84.48%	81.13%	71.14%
w/ HLAC	37.84%	85.20%	82.58%	86.58%
w/ LBP	94.21%	88.98%	95.79%	73.13%
w/ HOG	91.61%	74.68%	96.13%	64.31%

4.3. Evaluation of Acoustic Summary Statistics Extraction Scheme

We next experimentally validate the proposed summary statistics extraction. In the previous experiment, the TFR obtained by CQT transform had been proved to be superior for ASC task. In this test, we further investigate the feature weighting scheme for acoustic summary statistics extraction. Figure 9 presents the contribution of employing feature-enhancing weighting. It was evident that the proposed feature weighting scheme is universally applicable to generate efficient acoustic features for ASC. In addition, we further establish multiple spectro-temporal feature aggregation scheme to achieve superior ASC classification. Figure 10 presents the class-wise precision obtained by the proposed method. According to the results, our scheme produced ideal accuracies, i.e., over 98.5% accuracy, over the classes of bus, car, kid game, restaurant and high speed-train. While, there were several cases that current method failed in making a confident classification, such as for classifying the bust street, quiet street, and shop cases. In Table 2, we compared our result with state-of-the-art performances reported by the latest publications. It can be seen the proposed ASC scheme achieved superior accuracy compared to other methods. Notably, our approach also outperformed a very recent work using the optimal fusion of multiple convolution neural networks (CNNs) [24]. It is no doubt that CNNs is quite powerful learning method for general pattern classification tasks; however, in this study, the data size is small, and may induce under fitting of CNN models. That can be the main reason that our method can outperform CNNs. The comparison confirmed the superiority of the proposed scheme for ASC task.

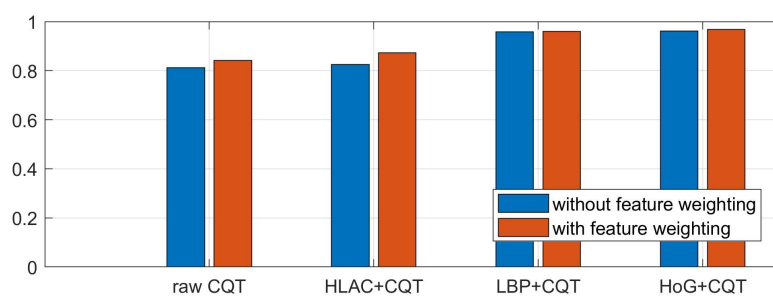
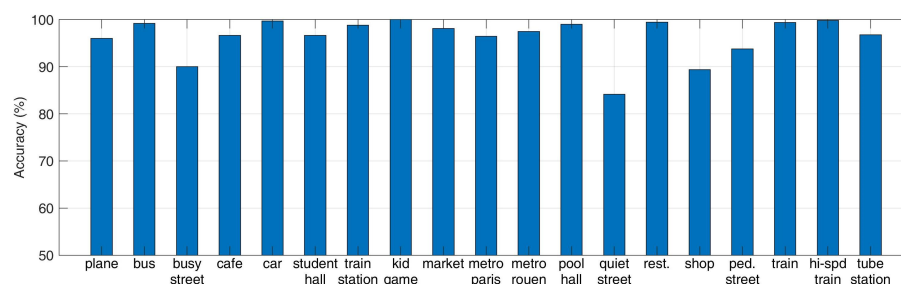
**Figure 9.** Classification accuracy comparison between w/o feature-enhancing weighting schemes.**Figure 10.** Classification accuracies for each scene category obtained by the proposed method.

Table 2. Performance comparison on the LITIS Rouen dataset.

	[14] CQT + HoG	[25] Spectral Weighting	[26] HoG + SPD	[13] Avg. Spectrum + BoW	[24] CNN	Proposed CQT + HoG + Weights	Proposed CQT + HoG + Weights + Fusion
Accuracy	91.14%	92.01%	93.4%	96.0%	96.6%	96.82%	96.98%

5. Conclusions

This paper presented a novel scheme for acoustic scene classification based on robust summary statistics extraction and efficient class score fusion. To characterize spectro-temporal patterns in sound, we evaluated various time-frequency representations with efficient local descriptors. Motivated by finding in psychological research—auditory system can keep relevant information about the acoustic environment, while weeding out the irrelevant details, We develop novel scheme to extract summary statistics for ASC to enhance discriminative scene-specific sound textures as well as to suppress scene-irrelevant events. Finally, we aggregate multi-way discriminant information characterized by various local descriptors through optimal weighted averaging. The proposed method is validated with Rouen dataset and experimental results presented superiority of proposed approach.

Author Contributions: J.Y. provided the method and conducted the experiments. T.K. was in charge of experiment design and parameter optimization. N.T., H.T. and M.M. supervised the theoretical statement of the problem. All authors revised the paper for intellectual content.

Funding: This research received no external funding

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chu, S.; Narayanan, S.; Jay Kuo, C.C. Environmental sound recognition with time-frequency audio features. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *17*, 1142–1158. [\[CrossRef\]](#)
2. Barchiesi, D.; Giannoulis, D.; Stowell, D.; Plumbly, M.D. Acoustic scene classification: Classifying environments from the sounds they produce. *IEEE Signal Process. Mag.* **2015**, *32*, 16–34. [\[CrossRef\]](#)
3. Wang, W. *Machine Audition: Principles, Algorithms and Systems*; IGI Global Press: Hershey, PA, USA, 2011.
4. McDermott, J.H.; Simoncelli, E.P. Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron* **2011**, *71*, 926–940. [\[CrossRef\]](#) [\[PubMed\]](#)
5. McDermott, J.H.; Schemitsch, M.; Simoncelli, E.P. Summary statistics in auditory perception. *Nat. Neurosci.* **2013**, *16*, 493–498. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Jrgen, T.; Geiger, B.S.; Rigoll, G. Recognising acoustic scenes with large-scale audio feature extraction and svm. *Tech. Rep.* **2013**.
7. Ellis, D.P.W.; Zeng, X.; Mcdermott, J.H. Classifying soundtracks with audio texture features. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011.
8. Salamon, J.; Jacoby, C.; Bello, J.P. A dataset and taxonomy for urban sound research. In Proceedings of the 22nd ACM international conference on Multimedia, Orlando, FL, USA, 3–7 November 2014.
9. Krijnders, J.D.; Ten Holt, G. A tone-fit feature representation for scene classification. *Energy* **2013**, *400*, 500.
10. Nelken, I.; de Cheveigne, A. An ear for statistics. *Nat. Neurosci.* **2013**, *16*, 381–382. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Yu, G.; Slotine, J.J. Audio classification from timefrequency texture. *arXiv* **2008**, arXiv:0809.4501.
12. Kobayashi, T.; Ye, J. Acoustic feature extraction by statistics based local binary pattern for environmental sound classification. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014.
13. Ye, J.; Kobayashi, T.; Murakawa, M.; Higuchi, T. Acoustic scene classification based on sound textures and events. In Proceedings of the 23rd ACM international conference on Multimedia, Brisbane, Australia, 26–30 October 2015.
14. Rakotomamonjy, A.; Gasso, G. Histogram of gradients of time-frequency representations for audio scene classification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 142–153.

15. Virtanen, T.; Mesaros, A.; Heittola, T.; Diment, A.; Vincent, E.; Benetos, E.; Elizalde, B. DCASE2017 Challenge Setup: Tasks, Datasets and Baseline System. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), Munich, Germany, 16–17 November 2017.
16. Hershey, S.; Chaudhuri, S.; Ellis, D.P.; Gemmeke, J.F.; Jansen, A.; Moore, R.C.; Plakal, M.; Platt, D.; Saurous, R.A.; Seybold, B.; et al. CNN architectures for large-scale audio classification. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017.
17. Valenti, M.; Squartini, S.; Diment, A.; Giambattista Parascandolo, G.; Virtanen, T. A convolutional neural network approach for acoustic scene classification. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017.
18. Mun, S.; Park, S.; Han, D.K.; Ko, H. Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane. In Proceedings of the Detection and Classification of Acoustic Scenes and Events, Munich, Germany, 16 November 2017.
19. Brown, J.C. Calculation of a constant q spectral transform. *J. Acoust. Soc. Am.* **1991**, *89*, 1. [[CrossRef](#)]
20. Shinohara, Y.; Otsu, N. Facial expression recognition using fisher weight maps. In Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition, Seoul, South Korea, 19 May 2004.
21. He, D.-C.; Wang, L. Texture unit, texture spectrum, and texture analysis. *IEEE Trans. Geosci. Remote Sens.* **1990**, *28*, 509–512.
22. Christopher, M.; Bishop, P.R.; Learning, M. *Information Science and Statistics*; Springer: Secaucus, NJ, USA, 2006.
23. Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
24. Phan, H.; Hertel, L.; Maass, M.; Koch, P.; Mazur, R.; Mertins, A. Improved audio scene classification based on label-tree embeddings and convolutional neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 1278–1290. [[CrossRef](#)]
25. Kobayashi, T.; Ye, J. Discriminatively learned filter bank for acoustic features. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016.
26. Bisot, V.; Essid, S.; Richard, G. HOG and subband power distribution image features for acoustic scene classification. In Proceedings of the 2015 23rd European Signal Processing Conference (EUSIPCO), Nice, France, 31 August–4 September 2015.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).