Acoustic Scene Classification based on Sound Textures and Events

Jiaxing Ye, Takumi Kobayashi, Masahiro Murakawa, Tetsuya Higuchi
National Institute of Advanced Industrial Science and Technology (AIST)
{jiaxing.you, takumi.kobayashi, m.murakawa, t-higuchi}@aist.go.jp

ABSTRACT

Semantic labelling of acoustic scenes has recently emerged as active topic covering a wide range of applications, e.g. surveillance and audio-based information retrieval. In this paper, we present an effective approach for acoustic scene classification through characterizing both background sound textures and acoustic events. The work takes inspiration from the psychoacoustic definition of acoustic scenes, that is, 'skeleton of (acoustic) events on a bed of (sound) texture'. In detail, we firstly employ distinct models to exploit sound textures and events in acoustic scenes, individually. Subsequently, based on fact that the perceptual importance of two parts will vary with respect to different scene categories, we develop favourable class-conditional fusion scheme to aggregate two-channel information. To validate proposed approach, we conduct extensive experiments on Rouen dataset which includes 19 categories of daily acoustic scenes with 3026 real-world recordings, and the proposed approach outperforms state-of-the-art methods by a large margin.

Categories and Subject Descriptors

H.2 [**Experience**]: Music, Speech and Audio Processing in Multimedia

Keywords

acoustic scene classification, constant-Q transform , Histogram of Gradients, model aggregation

1. INTRODUCTION

Motivated by various real-world applications, such as machine listening and surveillance, acoustic scene classification (ASC) has become a popular topic lately [1]. ASC aims to assign a semantic label to an input audio clip indicating the environment where it has been recorded. Unlike in conventional sound processing that speech and music are main objectives, ASC deals with comprehensive acoustics in daily life, such as sound of motor, footsteps and wind. Therefore, due to the complexity in environmental sound composition, conventional sound analysis schemes, such as the ones developed for automatic speech recognition and music information retrieval, are not well suited to ASC [1].

MM'15, October 26-30, 2015, Brisbane, Australia

DOI: http://dx.doi.org/10.1145/2733373.2806389

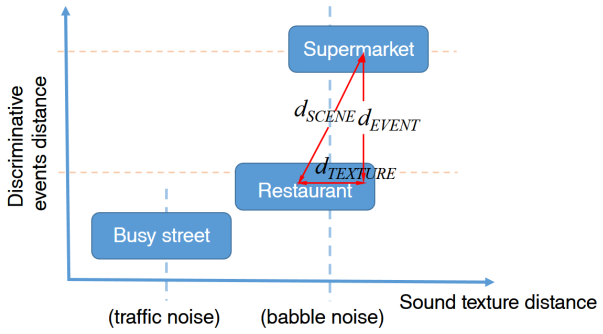


Figure 1: Illustration of conceptual psychoacoustic model for acoustic scene classification

ASC is quite a broad problem with all kinds of sounds in life involved. In order to design effective ASC scheme, it is critical to clarify the ASC task as well as to point out its distinction from other closely related problems, e.g. acoustic event detection (AED). In this regard, we introduce psychoacoustic definition of acoustic scenes, that is, "skeleton of (acoustic) events on a bed of (sound) texture" [7]. Clearly, there are two channels of information that contribute to acoustic scene perception, namely, sound textures which present high temporal homogeneity nature of environmental sound and acoustic events which are superimposed on the background. In Fig. 1 we present a conceptual chart for psychoacoustic model for ASC. And three major findings could be drawn from the chart:

† Perceptual distance between acoustic scenes should be obtained via characterizing both sound textures and key events [7].

† It is evident that AED is one sub-problem of ASC, which partially contributes to ASC analysis.

† The contributions for acoustic scene cognition of sound textures and events are anticipated to vary with respect to distinct scene classes.

Conventional studies on ASC mainly focus on sound texture information which is characterized by summary statistics over acoustic features. Various features have been evaluated for ASC, e.g. zero-cross rate (ZCR), MFCC and Gammatone filters [1, 10, 2]. Due to the high temporal homogeneity in sound textures, descriptive statistics, such as av-

erage and 1st/2nd order differential derivatives, are commonly adopted to build global textural feature [10]. Subsequently, various statistical models are employed to perform scene classification in a supervised manner. Compared with generative models like Gaussian mixture models (GMMs) and Hidden Markov model (HMM) [1], discriminative models, i.e. support vector machines (SVM), usually achieve superior results for ASC using multiple acoustic features with high dimension [4, 10]. It is noteworthy that bag-of-(acoustic)features (BoF) model has been regarded as standard model for ASC, however, latest investigation revealed its inefficiency in handling complex ASC with high within-class sound variability [5].

In this paper, we develop effective computational framework that approximates psychoacoustic auditory scene cognition to incorporate sound textures and events information for ASC. We firstly define general fusion rule of complementary information through convex combination as follows:

$$d_{SCENE} = \alpha_c \times d_E + (1 - \alpha_c) \times d_T,\ 0 \leq \alpha_c \leq 1\ , \quad (1)$$

where d_T and d_E denote perceptual distance for sound textures and acoustic events, respectively. α_c is the contribution weight with respect to certain sound scene class (indexed by c). Now the ASC task is decomposed into two sub-problems: sound texture analysis and acoustic event investigation. For the first sub-problem, we introduce effective acoustic features for sound texture extraction, subsequently probabilistic SVM (ProbSVM) is adopted to generate d_T in a supervised manner; for the latter sub-problem, BoF model is introduced to generate mid-level features representing acoustic events. Similarly, ProbSVM is applied again to the mid-level (event-based) features for estimating d_E. To fuse the two-way information , we adopt linear blending scheme on validation set to infer the optimal α_c. The main contributions of this paper can be outlined as follows:

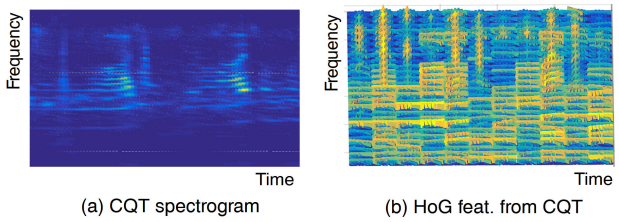
- We propose an effective computational framework that characterizes both sound textures and acoustic events for ASC. The proposed approach outperformed state-of-the-art results by a big margin.
- To incorporate sound textures and acoustic events information logically, we explore varying (relative) contributions of complementary components with respect to different scene categories.
- We introduce the effective acoustic features for characterizing environmental sounds and the superiority of the proposed feature has been validated through experimental comparisons.

2. THE PROPOSED APPROACH

In this section, we present the proposed approach in a top-down design. To approximate the acoustic scene perceptual distance with a statistical learning formulation, we introduce probabilistic estimates to rewrite (1) as:

$$S^c_{SCENE} = \alpha_c \times S^c_E + (1 - \alpha_c) \times S^c_T,\ 0 \leq \alpha_c \leq 1\ , \quad (2)$$

where S^c_T, S^c_E and S^c_{SCENE} denote class scores (membership probabilities) to c-category of scene which are derived from sound textures, acoustic events and both components, respectively. In this part, we explain how to estimate the scene-conditional fusion weights of α_c and the class scores of S^c_E, S^c_T as follows.



(a) CQT spectrogram (b) HoG feat. from CQT

Figure 2: Acoustic features demonstration

2.1 Acoustic feature extraction

To characterize wide spectro-temporal dynamics in environmental sounds, some recent works attempt to adopt image features, such as Local Binary Patterns (LBP) and Histogram of oriented Gradients (HoG), to extract acoustic feature for ASC in a two-dimensional manner and achieved favourable results [4, 9]. Some adjustments, such as pre- and post- processing, are commonly applied to make image descriptors better suit for ASC.

In a similar setting, we introduce acoustic features through extracting HoG patterns from Constant-Q transform (CQT) time-frequency representations (TFR). CQT increases time resolution towards higher frequencies, it is therefore preferred for describing environmental sounds [11]. To encode rich spectro-temporal patterns, we adopt HoG descriptors which effectively characterize 2-dimensional variations with cumulative oriented gradients over all local regions in spectrogram. In Fig.2, we show both CQT spectrogram (a) and HoG representation (b) of *cafe* scene, and from the charts, we see predominant spectro-temporal patterns are well retained by HoG. Meanwhile, the feature dimension of HoG is much lower compared to that of CQT spectrogram. Notably, acoustic features used in [9] is most close to ours, yet the key difference is, they added a preprocessing stage to resize CQT spectrogram to 512×512, hence discarding much detail time-frequency information. On the contrary, we directly apply HoG to CQT spectrogram and thus precise local information can be characterized for ASC.

2.2 Sound textures analysis

In this section, we present our path to characterize sound textures in acoustic scene.

2.2.1 *Sound textures feature extraction*

Recent study in neuroscience reveals that auditory system summarizes temporal details of sounds using time-averaged statistics for auditory scene cognition [6]. Grounded on such finding, we extract sound textures by performing average pooling along time on all acoustic feature vectors $[\mathbf{x}_1, ..., \mathbf{x}_N]$ from a sound scene:

$$\mathbf{x}_{texture} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i, \quad 1 \leq i \leq N\ , \quad (3)$$

where $\mathbf{x}_i$ is the ith (in time sequence) extracted HoG feature (from CQT spectrogram). $\mathbf{x}_{texture}$ denotes global sound textures of input acoustic scene.

2.2.2 *Class score estimation*

To estimate class score S^c_T in (2) based on textures information, we employ probabilistic SVM that estimates proba-

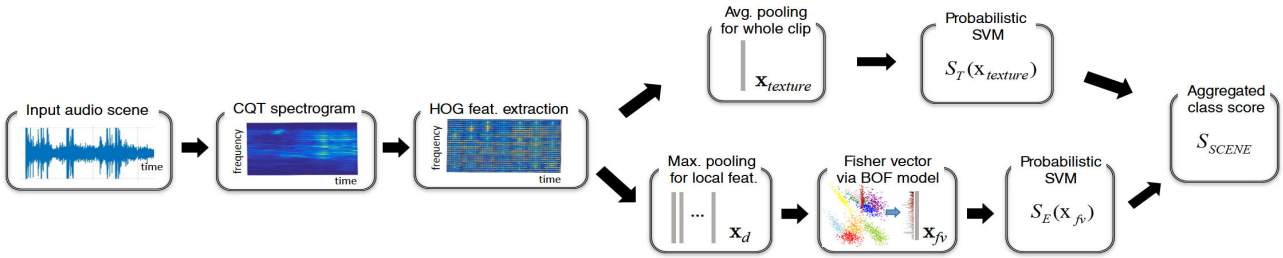


Figure 3: Overview of the proposed acoustic scene classification scheme

bility by investigating distance between input data and hyperplane in the (kernel) feature space. The formulation can be expressed as:

$$\min_{A,B} \frac{1}{M} \sum_{m=1}^{M} \log\left(1+\exp\left(-y_m(A(\mathbf{w}'_{svm}\Phi(\mathbf{x}_m)+b_{svm})+B)\right)\right) , \tag{4}$$

where $\{\mathbf{x}_m, y_m\}$ are training data and label. 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 $S^c_{texture}$ for input feature:

$$S_T(\mathbf{x}_{texture}) = \text{sigmoid}(A_T(\mathbf{w}'_T\Phi_T(\mathbf{x}_{texture})+b_T)+B_T) \tag{5}$$

where subscript T denotes parameters are derived from sound texture information.

2.3 Acoustic events analysis

Bag-of-feature (BoF) model is efficient to build (content-based) mid-level representation from low-level features. In sound processing, the BoF model has been widely studied for both ASC and AED [1, 8]. Latest studies revealed that BoF model is preferred for acoustic event detection [8], rather than for sound scene classification [5]. In this study, we employ BoF model to generate mid-level features that encode key events in auditory scenes.

(1) Low-level (dense) feature extraction. Given the waveform of acoustic signal, we firstly apply 1-second sliding window with half overlapping to segment data into slices. For each slice, we perform CQT transform and further extract HoG features. Finally, we perform max. pooling along time to enhance low-level feature for describing events and one vector can be obtained for each slice.

(2) Codebook generation. We employ Gaussian mixture model (GMM) to characterize distribution of input features:

$$p(\mathbf{x}_d;\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_d;\mu_k,\Sigma_k) \tag{6}$$

where $\mathbf{x}_d$ denotes acoustic feature, K is number of mixtures, and $\theta = \{\pi_k, \mu_k, \Sigma_k\}$ are parameters for GMM. Expectation maximization (EM) algorithm is applied to fit GMM model to given features.

(3) Encoding feature with codebook. Fisher vector (FV) delivers more robust performance for modelling patterns with noise [3], and hence, we employ FV to generate mid-level representation for acoustic events and it can be computed as follows:

$$\mathcal{G}^{\mathbf{x}_d}_{\mu,k} = \frac{1}{\sqrt{\pi_k}}\gamma_k\left(\frac{\mathbf{x}_d-\mu_k}{\sigma_k}\right), \ \mathcal{G}^{\mathbf{x}_d}_{\sigma,k} = \frac{1}{\sqrt{\pi_k}}\gamma_k\left[\frac{(\mathbf{x}_d-\mu_k)^2}{\sigma_k^2}-1\right] , \tag{7}$$

where γ_k is the weight of feature $\mathbf{x}_d$ to k-th Gaussian mixture. FV is obtained by concatenating these two gradients: $\mathbf{x}_{fv} = [\mathcal{G}^{\mathbf{x}}_{\mu,1}, \mathcal{G}^{\mathbf{x}}_{\sigma,1}, ..., \mathcal{G}^{\mathbf{x}}_{\mu,k}, \mathcal{G}^{\mathbf{x}}_{\sigma,k}]$, and then L_2 normalization is applied to facilitate classification. In the same vein as in 2.2.2, we fit ProbSVM model to FV features for estimating class score as follows:

$$S_E(\mathbf{x}_{fv}) = \text{sigmoid}(A_E(\mathbf{w}'_E\Phi_E(\mathbf{x}_{fv})+b_E)+B_E) , \tag{8}$$

where subscript E indicates sound event information.

2.4 Class score aggregation for ASC

Having computed class scores of $S_E(\mathbf{x}^{cj}_{fv})$ and $S_T(\mathbf{x}^{cj}_{texture})$, we further estimate optimal scene class-conditional fusion weights α_c via cross-validation. The objective function is defined as:

$$\min_{\alpha_c} \sum_{c=1}^{C}\sum_{j=1}^{N_c}\left(y_{cj}-(\alpha_c S_E(\mathbf{x}^{cj}_{fv})+(1-\alpha_c)S_T(\mathbf{x}^{cj}_{texture}))\right)^2 \tag{9}$$

where C is number of scene classes and cj denotes jth sample in cth class. α_c can be solved analytically, and overall class score S^c_{SCENE} can be computed through (2). Finally, class label is derived by $\text{argmin}_c(-S^c_{SCENE})$. To summarize whole process, a flow chart is presented in Fig. 3.

3. EXPERIMENTS

To validate the proposed scheme, we conduct experiments on Litis Rouen Dataset [9], which consists of 19 classes of real-world audio scenes with 3026 samples with 30 sec length. 20-fold splits are provided to partition data into 80%-training/20%-test sets. Within a training set, learning and validation sets are also prepared with 50%-50%.

In our experiments, we set number of bins per octave for CQT transform to be 48. In HoG extraction, local region size, number of orientations were set to 8×8 and 8, respectively. For codebook generation in BoF model, we used 256 GMM models (K = 256 in (6)) and Gaussian kernel was applied in ProbSVM.

3.1 Acoustic feature evaluation

In first test, we drew comparison between proposed features with Fourier transform(FT) and MFCC. In FT, we set analysis window to 30ms with half overlapping and for MFCC, 1st and 2nd order differential derivatives are also extracted and dimension was 39. We extracted sound textures

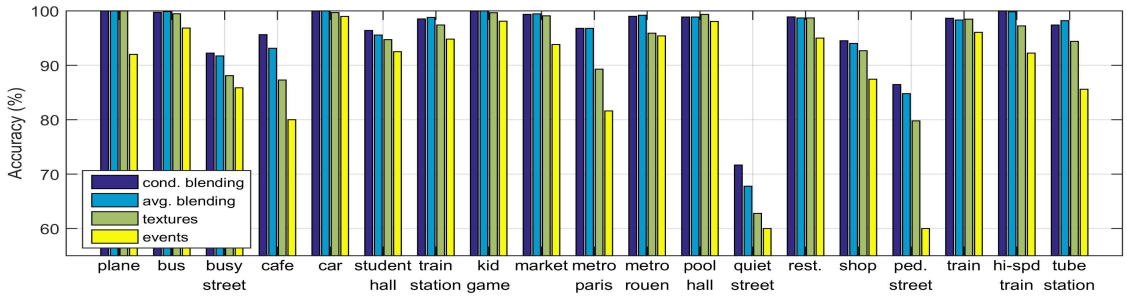


Figure 4: Classification experimental results comparison for whole Rouen dataset

Table 1: ASC accuracy for features test (%)

	MFCC	FFT	CQT
w/o HoG	70.9	81.84	88.68
w/ HoG	77.22	92.15	94.62

Table 2: Averaged ASC accuracy comparison (%)

Events	Textures	Avg. blend	Cond. blend	[9]'s
90.51	94.62	95.52	96.08	91.14

from all candidate features via time-averaging (3). Moreover, to clarify effectiveness of HoG features, we further compared cases of with/without HoG extraction. Results shown in Tab.1 brought us following conclusions: 1. CQT renders time-frequency representation that is well suited to ASC. 2. Over all three tests, HoG features contributed over 7.5% in accuracy on an average, hence its significance is evident.

3.2 Framework evaluation

We further conducted experiment to validate the proposed scheme, and results were summarized in Fig. 4 and Tab. 2. Three major conclusions can be drawn by examining the results: 1. Comparing to events, sound textures carry predominant scene information and thus exhibited higher classification accuracy. 2. Sound textures and events can compensate each other for achieving better ASC precision. 3. The proposed class-conditional fusion scheme had been validated and it always generated better results compared with average blending ($\alpha_c = 0.5$). Finally, the proposed approach achieved 96.08% average accuracy for Rouen dataset and outperformed previous method that presented average precision of 91.4% [9]

4. CONCLUSIONS

In this paper, we presented a novel framework for acoustic scene classification which imitates psychoacoustic auditory scene cognition process through characterizing both sound textures and events. Effective acoustic features are employed for describing sound textures and events, and we further incorporate the two-channel information with respect to their importance for scene classification. The framework achieved superior results in real data evaluation. Grounded on proposed formulation for ASC, various modifications can be made, such as extracting multi-resolution features and employing non-linear blending to fuse two-way information of textures and events. Those will be our future work.

5. ACKNOWLEDGEMENTS

This study was partly supported by Cross-ministerial Strategic Innovation Promotion Program (SIP) / Technologies for maintenance, renewal, and management for infrastructure, and supported by the New Energy and Industrial Technology Development Organization (NEDO), Japan.

6. REFERENCES

[1] D. Barchiesi, D. Giannoulis, D. Stowell, and M. Plumbley. Acoustic scene classification: Classifying environments from the sounds they produce. *Signal Processing Magazine, IEEE*, 32(3):16–34, May 2015.

[2] J. Chen, Y. Wang, and D. Wang. A feature study for classification-based speech separation at low signal-to-noise ratios. *IEEE TASLP*, 22(12):1993–2002, Dec. 2014.

[3] H. Jegou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3304–3311, June 2010.

[4] T. Kobayashi and J. Ye. Acoustic feature extraction by statistics based local binary pattern for environmental sound classification. In *IEEE ICASSP*, pages 3052–3056, May 2014.

[5] M. Lagrange, G. Lafay, B. Defreville, and J.-J. Aucouturier. The bag-of-frames approach: a not so sufficient model for urban soundscapes. *CoRR*, abs/1412.4052, 2014.

[6] J. H. McDermott, M. Schemitsch, and E. P. Simoncelli. Summary statistics in auditory perception. *Nature Neuroscience.*, 16:493 – 498, Apr. 2013.

[7] I. Nelken and A. de Cheveigne. An ear for statistics. *Nature Neuroscience.*, 16:381 – 382, Apr. 2013.

[8] A. Plinge, R. Grzeszick, and G. Fink. A bag-of-features approach to acoustic event detection. In *IEEE ICASSP*, pages 3704–3708, May 2014.

[9] A. Rakotomamonjy and G. Gasso. Histogram of gradients of time-frequency representations for audio scene classification. *IEEE TASLP*, 23(1):142–153, Jan 2015.

[10] J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research. In *22th ACM Int. Conf. on Multimedia*, Nov 2014.

[11] C. Schörkhuber and A. Klapuri. Constant-q transform toolbox for music processing. In *7th Sound and Music Computing Conf.*, July 2010.