# Discriminant Kernel Learning for Acoustic Scene Classification with Multiple Observations

Jiaxing Ye Takumi Kobayashi Hiroshi Tsuda Masahiro Murakawa

National Institute of Advanced Industrial Science and Technology
1-1-1 Umezono, Tsukuba, Ibaraki, Japan
(jiaxing.you, takumi.kobayashi, hiroshi-tsuda, m.masahiro)@aist.go.jp

***Abstract*—In this paper, we propose a novel kernel learning scheme for acoustic scene classification using multiple short-term observations. The method takes inspiration from the recent result of psychological research — "Humans use summary statistics to perceive auditory sequences"; we endeavor to devise computational framework imitating such important auditory mechanism for acoustic scene parsing. Conventional schemes usually encode spectro-temporal patterns with a compact feature vector by time-averaging, e.g. in Gaussian Mixture models (GMM). However, such integration may not be the ideal, since the arithmetic mean is vulnerable to extreme outliers which can be generated by sounds irrelevant to scene category. In this work, an effective scheme has been developed to exploit rich discriminant information from multiple short-term observations of an acoustic scene. Concretely, we first segment audio recording into short slices, e.g. 2 seconds; one vector can be extracted from each slice consisting of descriptive features. Then, we employ the resultant feature matrix to represent an acoustic scene. Since discriminant information of an acoustic scene can be characterized by either global structure or local patterns, we perform heterogeneous kernel analysis in hybrid feature spaces. Moreover, we conditionally fuse the two-way discriminant information to achieve better classification. The proposed method is validated using DCASE2016 challenge dataset. Experimental results demonstrated the effectiveness of our approach.**

## I. Introduction

Computational auditory scene analysis becomes more active area of research in recent years [1], [2]. It refers to the computational analysis of recording of acoustic scene, and the interpretation of useful information, such as location and interested activities. The recorded audio data, as multi-dimension stream fluctuating on time-frequency plane, usually embody high variations. To effectively perform acoustic scene recognition, it is preferred to systematically design the scheme considering the audio structure, similarity metric and classifier. This paper addresses the problem of acoustic scene classification (ASC) and proposes novel scheme to characterize rich discriminant information from acoustic scenes.

For decades, the research topics of Audio scene content analysis and retrieval have been long-standing [3]. Much research efforts have been spent on development of acoustic scene classification system using advanced signal processing and machine learning techniques [1], [4]. Although some progress has been made, the key issues in acoustic scene understanding, i.e. effective/robust feature representations and suitable framework for acoustic scene parsing, are still open questions to the research field.

In computational auditory scene analysis, one crucial issue is to extract efficient features to characterize acoustic scene [1], [5]. Standard approach to ASC firstly extracts time-frequency representations (TFRs) from audio signal, such as Mel-scale spectrogram and mel-frequency cepstral coefficients (MFCCs); then, statistical moments, including Gaussian mixture models (GMM), skewness and kurtosis, are employed to convert TFRs (matrix) to compact feature vector [6], [7], [8]. It is noteworthy that arithmetic mean played a key role throughout feature extraction.

Recent psychological studies reveal that "Humans use summary statistics to perceive auditory sequences" [9], [10]. Although aforementioned descriptive statistics can be employed as off-the-shelf tools to "summarize" acoustic scenes, human auditory system is functioning differently. More concretely, time-averaging statistics assume that every frame of audio data contains identical discrimination information for ASC. In contrast, auditory system adopts an adaptive scheme that "keep the relevant information about the environment, while weeding out the irrelevant detail", according to latest research report [11]. The mechanism had also been validated through listening test [12].

This paper attempts to achieve superior acoustic scene classification through establishing kernel discriminant analysis on feature matrix extracted from multiple observations of acoustic scene. In detail, we first segment audio clip into slices; then extract descriptive feature vector from every slice. A set of feature vectors can be obtained which is used to represent acoustic scene. Compared with conventional vector-wise features, proposed matrix representation retains plenty of local discriminant patterns. At classification stage, Gaussian mixture models (GMM) and support vector machines (SVM) are most extensively applied classifiers to perform vector-wise acoustic scene classification [1].[2]. Lately, inspired by the success of deep neural networks (DNN) in numerical application fields, e.g. computer vision and speech recognition applications, the method and its variants had been employed for content analysis of ambient sounds [2]. However, lack of large-scale labeled sound event data is the practical issue that deteriorates the performance of DNN-based ASC systems. In this study, in order to exploit discriminant information from acoustic fea-

ISBN 978-0-9928626-7-1 

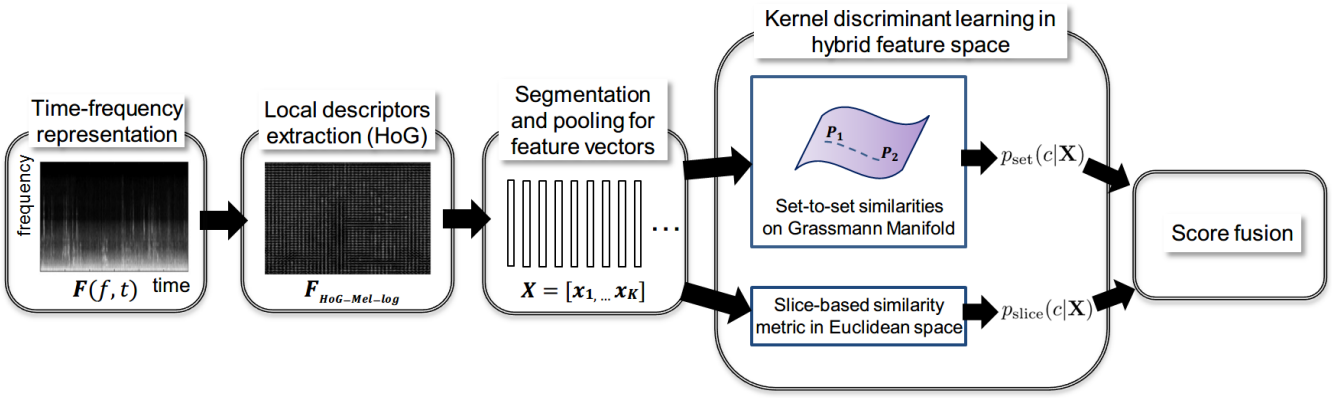


Fig. 1. Proposed acoustic scene classification scheme using heterogeneous kernel learning

ture matrix, we introduce heterogeneous similarity measures, including set-to-set Riemannian distance on Grassmann manifold and (column) instance-wise distance in Euclidean space. The hypothesis behind is that both short and long time observations can convey information to discern an acoustic scene. Moreover, according to result of latest DCASE2016 challenge [2], aggregation of multiple features/similarity metrics can greatly boost classification accuracy [13], [14], since it is possible way to integrate multiple-domain discriminant power for better classification. In a similar manner, we conditionally fuse scores obtained from different kernel feature spaces and perform classification on aggregated scores. In short, our major contributions lie in:

† We employ matrix-based acoustic scene feature representation for ASC task. In contrast to conventional features that rely on time-integrated feature vectors, the matrix-based feature retrains rich short-term spectro-temporal discriminant information which is anticipated to greatly contribute to classification task.

† Since discriminant information of an acoustic scene can be characterized by either global structure or local patterns, we introduce kernel methods to exploit discriminant information in hybrid feature spaces. In detail, heterogeneous kernel learning is adopted to investigate various notions of similarities, including both set-to-set similarity defined on Grassmann manifold (also a Riemannian space) and vector-wise similarity in Euclidean space. Furthermore, we effectively aggregate the scores obtained from multiple feature spaces to boost classification performance.

† Several hyperparameters, such as Fourier window length and slice length, play key roles in ASC [2]. We conducted extensive experiments on optimal hyperparameter selection. The results can facilitate further ASC research. Finally, the proposed framework has been validated with DCASE2016 Task 1 dataset and favorable performance has been achieved.

## II. Proposed Method

In this section, we describe our ASC approach with details. A schematic flowchart is shown in Fig. 1.

### A. Acoustic feature representation

*1) Short-term Fourier transform (STFT):* We first convert audio waveform to spectrogram by using short-time Fourier transform:

$$\mathbf{F}(f,t) = \sum_{n=0}^{N-1} s(n)w(n)e^{\frac{-j2\pi nf}{N}}, \quad (1)$$

where $s(n)$ denotes audio frame segmented by length $N$, $w(n)$ represents hamming window and the short time spectral column $F(f,t)$ at time $t$ can be derived. Fourier window length $N$ is crucial parameter in ASC, therefore we conduct experiments in Sec. 3 to choose optimal N. To reduce dimension of spectrogram, we employ 60-band mel-scale filter bank and logarithmic conversion is then applied on Mel frequency scale energies. The obtained time-frequency representation is denoted by $\mathbf{F}_{\mathrm{Mel-log}}(b,t)$, $b \in [1,60]$ is filter bank index.

*2) Histogram of oriented Gradients (HoG) local descriptor:* Latest research toward ASC manifests that 2-dimensional local descriptors are efficient for describing environmental sounds, such as using local Binary patterns (LBP) [15] and histograms of oriented gradients (HoG) [16]. In a similar vein, we adopt HoG descriptor to characterize spectro-temporal structures in acoustic scenes. The extracted time-frequency representation is expressed as $\mathbf{F}_{\mathrm{HoG-Mel-log}}$.

*3) Segmenting and pooling:* To extract concise feature (matrix) from acoustic scene, we perform segmentation and pooling over $\mathbf{F}_{\mathrm{HoG-Mel-log}}$. Firstly, we uniformly divide feature matrix into $K$ slices along time axis; then average- and max-pooling are carried out on each slice and two resultant vectors are concatenated to form acoustic feature. As a result, input audio waveform is converted to feature matrix denoted by $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_K] \in \mathbb{R}^{D \times K}$, where $\mathbf{x}_k \in \mathbb{R}^D$ represents acoustic feature vector from $k$-th slice.

### B. Kernel discriminant learning in hybrid feature space for ASC

Based on acoustic feature matrices, we employ heterogeneous kernel learning to perform pattern analysis in two

ISBN 978-0-9928626-7-1 © EURASIP 2017

feature subspaces, which are dedicated to characterize discriminant information from global and local structures. Then, resultant two-way posterior probabilities are aggregated so as to make acoustic scene classification.

*1) Kernel-based global (set-wise) similarity analysis:* In this section, we explain the set-to-set similarity measures employed for acoustic scene classification in kernel feature space. We start with a briefly review on background theory of grassmannian geometry and further introduce the kernel analysis scheme used in this study.

The acoustic scene feature matrix $\mathbf{X}$ can be represented by a linear subspace $\mathbf{P} \in \mathbb{R}^{D\times r}$ via eigenvalue decomposition:

$$\sum_{k=1}^{K} \mathbf{x}_k \mathbf{x}_k^\top = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^\top, \quad \mathbf{P}^\top\mathbf{P} = \mathbf{I}_r \tag{2}$$

where $\mathbf{P} = [\mathbf{p}_1, ..., \mathbf{p}_r]$ are eigenvectors and $\mathrm{diag}(\mathbf{\Lambda})$ are eigenvalues. All extracted orthonormal matrices describing acoustic scenes can be treated as a collection of linear subspaces, which are also the points on Grassmann manifold $Gr(D, r)$ (also a Riemannian manifold). The geodesic distance between two linear subspaces $\mathbf{P}_i$ and $\mathbf{P}_j$ in Euclidean space is defined as the principal angles, or canonical angles [17]. The standard approach to compute between-subspace angles is to apply SVD, in which

$$\begin{gathered}\mathbf{P}_i^\top \mathbf{P}_j = \mathbf{U}\mathbf{S}\mathbf{V}', \quad \text{where} \quad \mathbf{U}^\top\mathbf{U} = \mathbf{I}, \mathbf{V}^\top\mathbf{V} = \mathbf{I} \\ \mathbf{S} = \mathrm{diag}(\cos^2\theta_1, ..., \cos^2\theta_m)\end{gathered} \tag{3}$$

$\cos\theta_i$ is the cosine of the $i$th principal angle. $\cos\theta_1, ..., \cos\theta_d$, sorted in descend order, are known as canonical correlations. The principal angles $0 \leq \theta_1 \leq ..... \leq \theta_k \leq \pi/2$ can be computed from singular values.

Audio data of acoustic scenes are of a nonstationary and nonlinear nature. However, the principal angle metric can only deal with linear discriminant analysis in Euclidean space. As for non-linear classification, the Radial basis function (RBF) kernel has been proved effective for a variety of applications [1], which maps the data points to an infinite dimensional Hilbert space, where nearly-linear hyperplane can be found [18]. In the same vein, we employ kernel learning method on Grassmann manifold to enable non-linear classification of acoustic scenes with matrix feature representations. We first project similarity between two points $\mathbf{P}_i$ and $\mathbf{P}_j$ on Grassmann manifold to Euclidean space using positive definite Mercer kernels [19]. In this study, we use projection kernel, which can be expressed as

$$\mathcal{K}_{Pr} = ||\mathbf{P}_i^\top \mathbf{P}_j||_F^2, \tag{4}$$

the mapping corresponding to the kernel is given by $\Phi_{Pr}(\mathbf{P}) = \mathbf{P}\mathbf{P}^\top$. Various kernels can be further generated from $\mathcal{K}_{Pr}$, such as projection-RBF kernels:

$$\mathcal{K}_{Pr}^{ploy}(\mathbf{P}_i, \mathbf{P}_j) = (\gamma\mathcal{K}_{Pr}(\mathbf{P}_i, \mathbf{P}_j))^d, \tag{5}$$

and projection-polynomial kernels:

$$\mathcal{K}_{Pr}^{rbf}(\mathbf{P}_i, \mathbf{P}_j) = \exp(-\gamma||\Phi_{Pr}(\mathbf{P}_i) - \Phi_{Pr}(\mathbf{P}_j)||_F^2), \tag{6}$$

where $||\cdot||_F$ indicates Frobenius norm. Based on empirical studies, projection-RBF kernel achieved better classification performance on multiple tasks [19], and thus it is selected for our ASC approach.

Based on the Riemannian kernels, traditional learning methods operating in vector space can be exploited to classify data points (i.e. acoustic feature matrix) on the Riemannian manifolds for acoustic scene. For this work, we employ the LibSVM [20] implementation on our pre-calculated Riemannian kernel matrices for acoustic scene classification. Moreover, in order to perform multi-class acoustic scene classification with probability output, we adopt probabilistic SVM classifier which investigates distance between input data and hyperplane in the (kernel) feature space. One-versus-one scheme is adopted due to its superior multi-class classification performance [21]. As a result, we obtain posterior probability $p_{\mathrm{set}}(c|\mathbf{X})$ for input acoustic scene by performing Grassmannian kernel learning process demonstrated in this section.

*2) Kernel-based local (slice-wise) similarity analysis:* In above section, we show our path to analyze acoustic scenes using global information, which is conveyed by acoustic feature matrix. In addition to global patterns, spectro-temporal structures in short slice also contain rich discriminant information for ASC. To this end, we carry out scene classification on all acoustic vectors from sub-slices $[\mathbf{x}_1, ..., \mathbf{x}_K]$ using probabilistic SVM with RBF kernel. Based on posterior probabilities $[p(c|\mathbf{x}_1)), ..., p(c|\mathbf{x}_K))]$ generated from discriminant model, majority voting is conducted to produce scene-wise class score as follows.

$$\begin{gathered}p_{\mathrm{slice}}(c|\mathbf{X}) = \frac{1}{2} + \frac{(\sum_{k=1}^{K} p(c|\mathbf{x}_k)) - 1/2}{K} \\ \text{s.t.} \quad \mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_K]\end{gathered} \tag{7}$$

### C. Probability score fusion

To aggregate class scores generated from each feature space, we introduce weighted score fusion method. Let $c$ denote the scene category index. The fusion weight is denoted by $\lambda_c \in [0, 1]$. Final score can be computed as follows.

$$\mathrm{score}_c = \lambda_c \times p_{\mathrm{set}}(c|\mathbf{X}) + (1 - \lambda_c) \times p_{\mathrm{slice}}(c|\mathbf{X}). \tag{8}$$

The convex weighting factor $\lambda_c$, which governs contributions of two-level discriminant information distilled from both global and local observations, can be empirically estimated using validation set.

## III. Experimental results

In this part, we present experimental validation of proposed approach on real-world data.

### A. Dataset and parameter settings

We validate proposed scheme using DCASE2016 Challenge Task 1 dataset, which contains 15 classes of acoustic scenes. The length of recording is 30 second, sampling frequency and bit depth are set to 44.1 kHz and 16 bits, respectively. In validation set, each class has 78 audio segments. The

ISBN 978-0-9928626-7-1 © EURASIP 2017

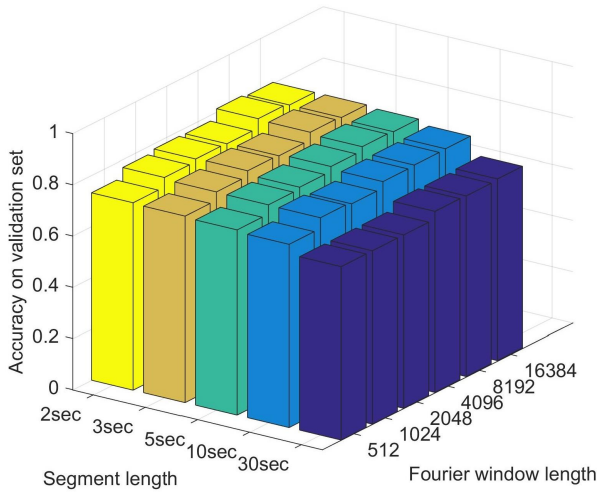


Fig. 2. ASC accuracy with different Fourier window and segmentation length

TABLE I
COMPOSITIONS OF ASC SYSTEMS IN COMPARISON

| | Feature | Classifier |
|---|---|---|
| Method 1 | Averaged Mel-log-spec. | RBF SVM |
| Method 2 | Averaged HoG-Mel-log-spec. | RBF SVM |
| Method 3 | Segmented HoG-Mel-log-spec. | RBF SVM + voting |
| Method 4 | Segmented HoG-Mel-log-spec. | Proposed HKL |

indices for train/validation split are provided. The evaluation set includes 390 clips. The parameter $\gamma$ in projection-RBF Riemannian kernel was set to 1.2. Gaussian scaling parameter and C regularization balancing weight in segment-wise SVM classifier were set to 0.5 and 3, respectively.

### *B. Hyperparameters tuning*

Recent articles revealed that several hyperparameters used for acoustic feature extraction played key role for ASC [2], [22], i.e. Fourier window length $N$ and segmentation length denoted by $L$ ($L = 30/K$). In pursuit of optimal hyperparameter combination, we conducted grid search on each pair $(N, K)$ to examine classification performance on validation set. Simple time-averaged Mel-log-spectrogram features were applied with linear SVM classifier in the test. Overall accuracies correspond to hyperparameter combinations are shown in Fig. 2. It can be seen that best accuracy reached to 81.20% by choosing $N = 8192, L = 2$. These settings were applied in following experiments. We hope the hyperparameter tuning results can facilitate further research for ASC.

### *C. Results on validation set*

To validated proposed approach, we perform experiments using DCASE2016 challenge Task 1 dataset. First, experimental comparison was drawn between four methods on validation set. The four tests were designated to testify the contributions of major components in proposed system and we showed their characteristics in Tab. 1. The ASC results

TABLE II
FOLDER-WISE ASC ACCURACIES COMPARISONS ON VALIDATION SET (%)

| | Method 1 | Method 2 | Method 3 | Method 4 |
|---|---|---|---|---|
| f1 | 73.10 | 75.17 | 82.76 | **88.97** |
| f2 | 59.66 | 68.62 | 82.41 | **86.21** |
| f3 | 72.48 | 74.16 | 79.53 | **85.91** |
| f4 | 76.71 | 74.66 | 80.14 | **84.59** |

TABLE III
CLASS-WISE ASC ACCURACIES COMPARISONS ON VALIDATION SET (%)

| | Method 1 | Method 2 | Method 3 | Method 4 |
|---|---|---|---|---|
| Beach | 69.2 | 78.2 | 87.2 | 93.6 |
| Bus | 79.5 | 84.6 | 93.6 | 97.4 |
| Cafe | 52.6 | 53.9 | 83.3 | 87.2 |
| Car | 83.3 | 87.2 | 94.9 | 94.9 |
| City | 89.7 | 78.2 | 89.7 | 91.0 |
| Forest | 89.7 | 82.1 | 92.3 | 93.6 |
| Grocery | 89.7 | 84.6 | 89.7 | 91.0 |
| Home | 64.1 | 85.9 | 79.5 | 91.0 |
| Library | 66.7 | 67.9 | 85.9 | 88.5 |
| Metro | 74.4 | 79.5 | 87.2 | 92.3 |
| Office | 76.9 | 75.6 | 73.1 | 91.0 |
| Park | 35.9 | 56.4 | 68.0 | 73.1 |
| Resident | 61.5 | 58.9 | 61.5 | 68.0 |
| Train | 44.9 | 38.5 | 42.3 | 53.9 |
| Tram | 79.5 | 85.9 | 89.7 | 89.7 |

generated by using the four methods were demonstrated in Tab. 2 and Tab. 3, in terms of folder-wise and category-wise performance, respectively. By comparing results made by first and second methods, we confirmed effectiveness of HoG features for local spectro-temporal characterization. Significant performance improvement can be seen from third column, in which scheme of classification using frame-averaged acoustic features were replaced by slicing / majority voting manner. Finally, by examining the 4-th column, proposed approach achieved superior classification precision in all 15 scene categories. Experimental comparisons proved the effectiveness of proposed heterogeneous kernel learning approach which improved ASC performance with large margin in both folder-wise and class-wise evaluation.

### *D. Results on evaluation set*

We further evaluated proposed scheme on evaluation set. The optimal parameters estimated at validation stage were adopted for the test. In a similar vein to previous test, experimental comparison was drawn among four methods. Tab. 4 presented classification accuracies. Besides, in order to facilitate detailed comparison with other studies, we show the confusion matrix in Fig. 3. As a result, we achieved overall accuracy of 88.97%. We obtained 100% classification for four scene classes, which are Bus, Forest path, Office and Tram. The worse accuracy were obtained for three classes,

TABLE IV
ASC ACCURACIES COMPARISONS ON EVALUATION SET (%)

| Method 1 | Method 2 | Method 3 | Method 4 |
|---|---|---|---|
| 80.0 | 82.05 | 85.38 | **88.97** |

ISBN 978-0-9928626-7-1 © EURASIP 2017

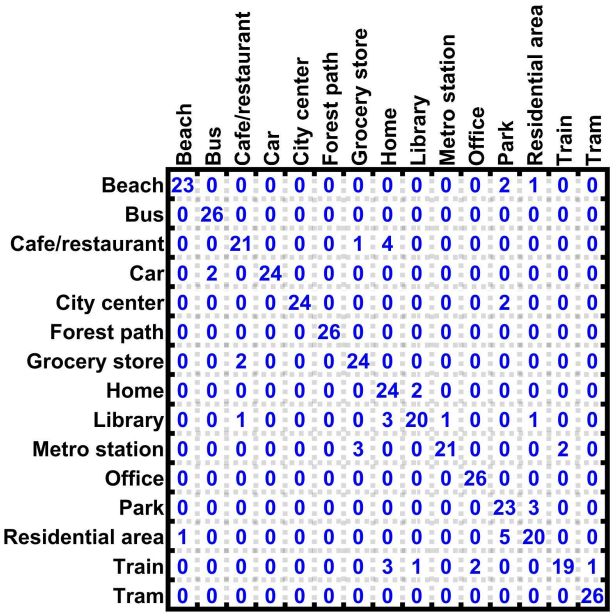


Fig. 3. ASC confusion matrix on evaluation set

which are library (76.92%), residential area (76.92%) and train (73.08%). Class-wise recognition performance is consistent to that from validation data. Compared to other submissions to DCASE2016 challenge Taks 1, Our system exhibit lower variance in terms of classification accuracy among 15 scene classes [2]. It is noteworthy that proposed method obtained 76.92 % precision which outperformed the best results (69.2%) reported in DCASE2016 challenge.

## IV. Conclusion

In this paper, a heterogeneous kernel-based learning approach for acoustic scene classification had been proposed which can effectively characterize discriminant in audio data. Unlike conventional methods which adopt single feature vector to represent acoustic scene, we employ a matrix consists of multiple acoustic feature vectors extracted from slices. Furthermore, we investigate set-to-set similarities in multiple kernel feature spaces which are anticipated to be well-suited to acoustic scene classification task, such as between-subspace principal angle metric defined in Euclidean space and Riemannian distance on Grassmann manifold. Besides, in order to incorporate discriminant information among multiple feature spaces, we employed optimal fusion rule for better classification. To validate proposed approach, we carried out extensive experiments on DCASE2016 challenge Task 1 dataset. The test results demonstrated through comparisons with other methods. In addition, the proposed approach can be further incorporated with multiple kernel learning and metric learning frameworks and those will be left to our future works.

## References


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

[2] Tuomas Virtanen, Annamaria Mesaros, Toni Heittola, Mark D. Plumbley, Peter Foster, Emmanouil Benetos, and Mathieu Lagrange, *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, Tampere University of Technology. Department of Signal Processing, 2016.

[3] Wang Wenwu, "Machine audition: Principles, algorithms and systems," *IGI Global Press*, 2011.

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

[5] Jrgen T. Geiger, Bjoern Schuller, and Gerhard Rigoll, "Recognising acoustic scenes with large-scale audio feature extraction and svm," Tech. Rep., 2013.

[6] Daniel P. W. Ellis, Xiaohong Zeng, and Josh H. Mcdermott, "Classifying soundtracks with audio texture features," in *Proc. ICASSP, to appear, Prague, 2011. Soundtrack Classification - Dan Ellis*.

[7] 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.

[8] Johannes D Krijnders and Gineke Ten Holt, "A tone-fit feature representation for scene classification," .

[9] Josh H. McDermott and Eero P. Simoncelli, "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis," *Neuron*, vol. 71, no. 5, pp. 926 – 940, 2011.

[10] Josh H. McDermott, Michael Schemitsch, and Eero P. Simoncelli, "Summary statistics in auditory perception," *Nature Neuroscience.*, vol. 16, pp. 493 – 498, Apr. 2013.

[11] Israel Nelken and Alain de Cheveigne, "An ear for statistics," *Nature Neuroscience.*, vol. 16, pp. 381 – 382, Apr. 2013.

[12] Vesa TK Peltonen, Antti J Eronen, Mikko P Parviainen, and Anssi P Klapuri, "Recognition of everyday auditory scenes: potentials, latencies and cues," .

[13] Sangwook Park, Seongkyu Mun, Younglo Lee, and Hanseok Ko, "Score fusion of classification systems for acoustic scene classification," .

[14] Anurag Kumar, Benjamin Elizalde, Ankit Shah, Rohan Badlani, Emmanuel Vincent, Bhiksha Raj, and Ian Lane, "DCASE challenge task 1," Tech. Rep., DCASE2016 Challenge, September 2016.

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

[16] Jiaxing Ye, T. Kobayashi, M. Murakawa, and T. Higuchi, "Acoustic scene classification based on sound textures and events," in *23th ACM Int. Conf. on Multimedia*, Oct 2015.

[17] Jihun Hamm and Daniel D Lee, "Grassmann discriminant analysis: a unifying view on subspace-based learning," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 376–383.

[18] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[19] Mehrtash T Harandi, Mathieu Salzmann, Sadeep Jayasumana, Richard Hartley, and Hongdong Li, "Expanding the family of grassmannian kernels: An embedding perspective," in *European Conference on Computer Vision*. Springer, 2014, pp. 408–423.

[20] Chih-Chung Chang and Chih-Jen Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27, 2011.

[21] Chih-Wei Hsu and Chih-Jen Lin, "A comparison of methods for multiclass support vector machines," *IEEE transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[22] Gustavo Mafra, Ngoc Duong, Alexey Ozerov, and Patrick Pérez, "Acoustic scene classification: An evaluation of an extremely compact feature representation," in *Detection and Classification of Acoustic Scenes and Events 2016*, 2016.


ISBN 978-0-9928626-7-1 © EURASIP 2017