KERNEL DISCRIMINANT ANALYSIS FOR ENVIRONMENTAL SOUND RECOGNITION BASED ON ACOUSTIC SUBSPACE

Jiaxing Ye Takumi Kobayashi Masahiro Murakawa Tetsuya Higuchi

National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan
{jiaxing.you, takumi.kobayashi, m.murakawa, t-higuchi}@aist.go.jp

ABSTRACT

In this paper, we propose an effective discriminant subspace learning framework to recognize the environmental sounds. Firstly, Gabor transform is adopted to characterize the time-frequency distributions of environmental sounds. We further encode the prominent time-frequency patterns with low rank representation by extracting the subspace from Gabor spectrogram. Unlike conventional sound recognition schemes that are mostly based on acoustic feature *vectors*, we treat the acoustic *subspaces* (matrixes) as basic elements for recognition, retaining rich temporal-spectral contextual information. At recognition stage, we employ kernel Fisher discriminant analysis to effectively exploit the class conditional distributions of environmental sounds which are favorable for performing multi-class classification. With a well developed kernel function, the proposed approach achieved superior recognition performance on RWCP sound scene database, compared with the existing methods.

Index Terms— Environmental sound, Gabor transform, subspace learning, kernel fisher discriminant analysis, canonical angle, RWCP Sound Scene Database

1. INTRODUCTION

Speech and music are the most informative sounds for human perception and have been extensively investigated for more than half a century. On the other hand, in real-world auditory scenes, apart from speech and music, there are various non-speech sounds carrying prominent information for human perception, such as phone bell, the glasses sounds and even alarms. These environmental acoustic events play important role in awareness of auditory context and could be widely applied in human-machine interaction, robotics and surveillance, etc.

The goal of environmental sound analysis is to retrieve significant information from auditory scenes. In recent years, more studies have been issued to challenge the environmental sound recognition task [1]. Based on the studies on speech recognition, modern environmental sound recognition systems are basically formulated into two stages: feature extraction and classification. At acoustic feature extraction stage, due to the wide non-stationary time-frequency variations, conventional speech features, such as Mel-frequency cepstral coefficients (MFCCs), are no longer favorable for characterizing environmental sounds. Various advanced signal processing techniques have been examined, such as Gabor and wavelet transforms [2]. Recently, Matching Pursuit (MP) [3] and non-negative matrix factorization (NMF) [4], the techniques that produce the sets of basis to form approximate signal representations, have been introduced for environmental sounds feature extraction and achieved some improvements in recognition performance. At recognition stage, conventional classifiers, such as Gaussian mixture models (GMM) [5], artificial neural networks (ANN) [6] and support vector machine (SVM) [7] have been evaluated for classifying environmental sounds. Some recent research revealed hidden Markov models (HMM) with multiple hidden states presents higher recognition accuracy than other classifiers [8]. There are also some proposals integrated several classification methods to make further improvement [9].

This study addresses the content based recognition on environmental sounds. Based on the survey above, we outline the contributions of this work by comparing with previous works:

1. We formulate the environmental sound recognition problem by subspace learning. The acoustic subspaces (matrixes) are treated as basic units for classification, thus the rich time-frequency inter-frame contextual information can be portrayed by subspace representation. Conversely, conventional sound recognition systems always employ the vector-based schemes. The feature extraction procedures are mainly carried out within the analysis frame and such scheme leads to deterioration in inter-frame temporal-spectral contextual information, such as in MFCCs. Although the first and second order derivations of MFCCs over time can be adopted to describe the inter-frame dynamics, the information loss at feature extraction stage cannot be fully compensated.

2. We adopt kernel Fisher discriminant analysis (KFDA) [10] for environmental sound recognition. Kernel methods have been extensively studied in machine learning and computer vision fields for decades due to their excellent performance in non-linear classification. For environmental

sound recognition, the most applied method is SVM, which addresses separating the samples through building optimal non-linear hyper plane corresponding to the labels of the data. Nevertheless, the class conditional distributions are ignored by SVM, which are preferred for multi-class classification. We therefore introduce KFDA scheme to effectively exploit the class conditional distributions to produce the optimal projection kernel feature space for conducting multi-class environmental sound recognition. Particularly, as for dealing with the acoustic subspace (matrix), we develop mutual subspace canonical kernel to effectively explore the distances between environmental sounds. The proposed discriminant analysis framework and kernel function performed well in experimental evaluation.

2. PROPOSED ENVIRONMENTAL SOUND RECOGNITION SCHEME

There are two major components in the proposed kernel discriminant recognition framework. The first is the feature extraction phase in which the audio data is transformed into subspace representation. Subsequently, at discriminant classification stage, we employ KFDA analysis to characterize the class conditional distribution and map the data into kernel discriminant feature space which is favorable for performing classification. Fig. 1 shows the schematic diagram of the proposed approach. Each step will be explained explicitly in this section.

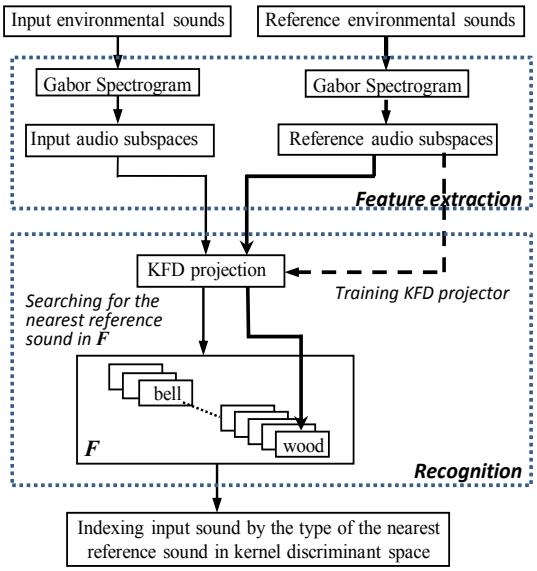


Fig. 1. Chart flow of the proposed framework

2.1. Environmental Sound Feature Extraction

Effective feature extraction is determinative to realize fine recognition performance. In here, we explain our feature extraction procedure and clarify the considerations.

Comparing with speech, environmental sounds manifest two fundamental differences: (*1*) Speech presents formant structures while environmental sounds do not exhibit such characteristics. (2) Speech is stationary signal with stable frequency distribution in short time, whereas most environmental sounds are non-stationary with wide temporal-spectral dynamics. Based on these two properties, we select Gabor transform for to characterize the non-environmental sounds.

Gabor transform is an effective time-frequency analysis tool for investigating acoustic signal which can be expressed as:

$$G\{x(n)\}_{(\tau,\omega)} = \sum_{n=0}^{N-1} x[n]\mathrm{w}[n-\tau]e^{-j\omega n}. \qquad (1)$$

where $x[n]$ is the signal to be analyzed, $\mathrm{w}[n]$ denotes the Gaussian window function over the framed signal with length N. Comparing with short-time Fourier transform, Gabor transform enables best simultaneous resolution in both time and frequency domains.

Subsequently, we encode the Gabor spectrogram with subspace representation. Low-rank subspaces can empirically approximate the structural distribution of the data as well as the variations, hence has been successfully applied in computer vision field [11], i.e. to express a set of images of face under varying lighting conditions and poses. In this study, we adopt subspace to describe acoustic signal based on two motivations: first, subspace representation effectively characterizes the prominent temporal-spectral distributions in audio data and discards the minor patterns, which are mostly noises in sounds, simultaneously; second, benefited from the much lower feature dimension, processing acoustic subspace is more efficient compared to dealing with raw feature vectors (Gabor spectrogram). We explain the procedure of extracting acoustic subspace as follows.

Let $G=[g_1, g_2, \cdots, g_T]$, $g_t(t=1,\cdots,T)\in\mathbb{R}^F$ denote Gabor spectrogram, t denotes the frame indexes and F stands for frequency coordinate. To formulate the Gabor spectrogram with subspace representation, we calculate eigenvalues $\Lambda = diag(\lambda_1,\ldots,\lambda_F)$ and eigenvectors $U = [u_1,\ldots,u_F]$ by:

$$R_{Cov_G}U = U\Lambda, \quad R_{Cov_G} \triangleq \sum_{t=1}^{T} E\{g_t g_t'\}, \qquad (2)$$

where $g_t', t\in(1,\cdots,T)$ is the transpose of g_t.

$U = [u_1,\ldots,u_F]$ span the acoustic subspace characterizing the time-frequency distributions of sound. The contribution ratio of eigenvector u_f in U is defined as:

$$\eta_f \triangleq \lambda_f \Big/ \sum_{i=1}^{F}\lambda_i, \qquad (3)$$

which denotes significance of corresponding eigenvector for expressing the audio data. The eigenvectors are ranked by their contribution ratios in the decreasing order. We can select the first K principle eigen vectors with higher contribution ratios and employ the subspace $U_K = [u_1,\ldots,u_K]$, $1 < K < F$ to express the main acoustic patterns. In addition,

the principle eigenvectors in U_K are normalized by their contribution ratios through $\tilde{u}_k = \lambda_k \Big/ \sum_{i=1}^{K} \lambda_i \cdot u_k, k = 1,\ldots,K,$ in a similar manner to [12]. The contribution weightings give prominence to the principle eigenvectors in describing audio data. Based on the procedures explained in this section, we extract acoustic subspaces from sounds.

2.2. Kernel Fisher Discriminant Analysis

In this section, we introduce the kernel Fisher discriminant analysis [10] that effectively characterizes class conditional distributions for environmental sound recognition.

Let $X_i = \{x_1^i, \ldots x_{N_i}^i\}$ be audio samples from class *i*. $\Phi(x)$ is the nonlinear mapping of the input vector *x* into kernel feature space $\mathcal{F}$. Kernel Fisher discriminant seeks the direction $w \in \mathcal{F}$ maximizing the Rayleigh quotient as:

$$J(w) = \frac{w^T S_B^\Phi w}{w^T S_W^\Phi w}, \tag{4}$$

$$\text{with } S_B^\Phi = (m_1^\Phi - m_2^\Phi)(m_1^\Phi - m_2^\Phi)^T \tag{5}$$

$$\text{and } S_W^\Phi = \sum_{i=1,2} \sum_{x \in X_i} (\Phi(x) - m_i^\Phi)(\Phi(x) - m_i^\Phi)^T, \tag{6}$$

$$m_i^\Phi = \frac{1}{N_i} \sum_{n=1}^{N_i} \Phi(x_n^i), \tag{7}$$

Because the solution is determined by scalar products only, instead of mapping the data explicitly to $\mathcal{F}$, the kernel trick is employed to compute these dot products, i.e. $k(x,y) = (\Phi(x) \cdot \Phi(x))$. Many kernel functions have been developed, the most applied one is the Gauss kernel $k(x,y) = \exp(-\|x-y\|^2 / \sigma)$. Then, the *w* can be represented by the linear combination of training samples as:

$$w = \sum_{n=1}^{N} a_n \Phi(x_n). \tag{8}$$

Consider the formula in expansion (4), we have

$$w^T m_i^\Phi = \frac{1}{N_i} \sum_{n=1}^{N} \sum_{j=1}^{N_i} a_n k(x_n, x_j^i) = a^T \mathbf{M}_i \tag{9}$$

where we express $(\mathbf{M}_i)_n = 1/N_i \sum_{j=1}^{N_i} k(x_n, x_j^i)$ by replacing the dot products by kernel function. Based on (9), the representation in (4) can be rewritten as

$$J(w) = \frac{w^T S_B^\Phi w}{w^T S_W^\Phi w} = \frac{a^T \mathbf{M} a}{a^T \mathbf{N} a} \tag{10}$$

where $\mathbf{M} = (\mathbf{M}_1 - \mathbf{M}_2)(\mathbf{M}_1 - \mathbf{M}_2)^T$, $\mathbf{N} = \sum_{i=1,2} K_i (\mathbf{I} - \mathbf{1}_{N_i}) K_i^T$ and K_i is the $N \times N_i$ kernel matrix for class *i* with $(K_i)_{nj} = k(x_n, x_j^i)$. $\mathbf{I}$ is identity matrix and $\mathbf{1}_{Ni}$ is the matrix with all $1/N_i$ entries. The maximum *J* can be obtained by $a = \mathbf{N}^{-1}(\mathbf{M}_1 - \mathbf{M}_2)$. The projection of input feature *x* onto *w* is given by

$$(w, \Phi(x)) = \sum_{n=1}^{N} a_n k(x_n, x). \tag{11}$$

Thus, we obtain the projected data without mapping to $\mathcal{F}$.

Kernel function determines the mapping feature space, which is critical for making classification. In this study, we treat the acoustic subspaces as basic elements. That means x_n^i is matrix, i.e. with $f \times k$ dimension. To investigate the subspace distance, we develop mutual subspace canonical kernel. We exhibit the details as follows.

Let two acoustic subspaces denoted by $x, y \in \mathbb{R}^{f \times k}$. We employ canonical angles to measure the distances between them [13], which is defined by:

$$\begin{aligned} &\cos\theta_p = \max_{u_p \in x} \max_{v_p \in y} u_p' v_p \qquad subject\ \ to \\ &u_p' u_p = 1, \quad u_p' u_q = 0, \\ &v_p' v_p = 1, \quad v_p' v_q = 0, \ (p = 1,\ldots,q-1) \end{aligned} \tag{12}$$

where u_p is the eigenvector of *x* and v_p is the eigenvector of y. $0 \le \theta_1 \le \cdots \le \theta_p \le \pi/2$ are the canonical angles between two subspaces. We adopt the minimum canonical angle for measuring the difference between the acoustic subspaces, which is defined as:

$$\min\_angle(x,y) = \theta_1 \tag{13}$$

Then, the proposed mutual subspace canonical kernel can be written as:

$$k(x,y) = \exp(-\|\min\_angle(x,y)\|^2 / \sigma) \tag{14}$$

Based on the proposed kernel function and discriminant learning scheme, the class conditional distribution can be effectively exploited. Based on the projected data representation, we employ nearest neighbor (NN) scheme with cosine distance to classify the input sound in the kernel discriminant feature space.

3. EXPERIMENTAL RESULTS AND DISCUSSION

To evaluate the proposed framework, we conduct environmental sound recognition experiments using real-world database. In this section, we demonstrate the experimental validation procedure and present the results.

3.1. Experimental Setup

3.1.1. Dataset

We employ Real World Computing partnership's (RWCP) sound scene database [14] to evaluate the proposed scheme. The RWCP database includes 105 types of environmental sound generated by 3 categories of sound sources: collision, action and characteristic sound sources. There are 9722 recording samples with the length ranging from about 1 second to several seconds. The sound clips were recorded with 48 kHz sampling rate and 16bit resolution.

3.1.2. Parameters selection

In the proposed framework, there are some parameters to be settled in advance. The short-time analysis window was fixed to 10ms with half size overlapped. We examined the contribution ratios of eigenvectors in acoustic subspace by

function (3) to determine the subspace dimension. Fig. 2 illustrated the contribution ratio distributions over eigenvectors of 4 types of environmental sounds. The chart manifested the acoustic characteristic can be effectively conveyed by first several eigenvectors. The acoustic subspace dimension selection was investigated based on experiments. The spread parameter σ in kernel function was set empirically to 3.4.

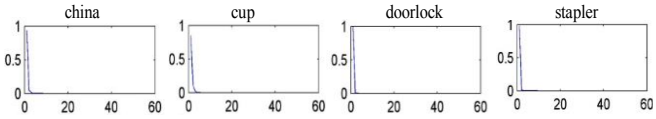


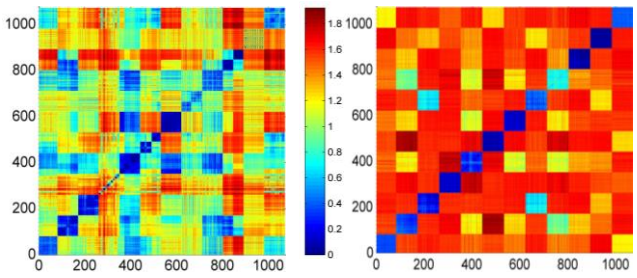
Fig. 2. Contribution ratios of first 60 eigenvectors in the subspaces of 4 kinds of environmental sounds

3.2. Experiments and Results

Two experiments have been conducted. The first was in the interest of manifesting the separability of the proposed approach. Subsequently, we validated the proposed scheme over full RWCP sound scene database with large amount of environmental audio data.

3.2.1. Validation of sound class separability of the proposed discriminant analysis scheme

Distinct between-class distances along with low within-class distances are demanded for achieving favorable recognition performance, which is presented by separability. We selected 12 classes of environmental sounds from RWCP database to verify the separability of the proposed scheme. The selected sounds consisted of impulsive sounds of wood, book, metal, glass cup, coins, hands clapping, dices, drum, doorlock and the sustaining sounds of particle dropping, spray and phone beeping, which are the same as the data used in [8]. Each sound class included 100 samples. We conducted 10-folder cross-validation on the selected dataset. For each type of sound, 90 samples were used for training and the remaining 10 clips were for testing. Fig.3 portrayed the pairwise cosine distance matrixes of training sounds in original feature space and kernel discriminant space. The gain in class separability by using the proposed framework can be clearly observed. The within-class distances depicted by 90×90 diagonal blocks are very small (in cold color) and the between-class distances on all other positions are much bigger (in warm colors) by contrast. We examined the recognition rates corresponding to the sound subspace dimension variations. The results in Fig. 4 manifested the highest recognition rate of 99.16% was achieved by adopting the first 2 principle eigenvectors to construct the sound subspace. The proposed approach outperformed the method in [8], which presented a recognition rate about 93% by using Matching Pursuit over Gabor features and HMM classifier on same dataset.



(1) Primitive feature space (2) kernel discriminant space

Fig. 3. Distance matrixes comparison of training data

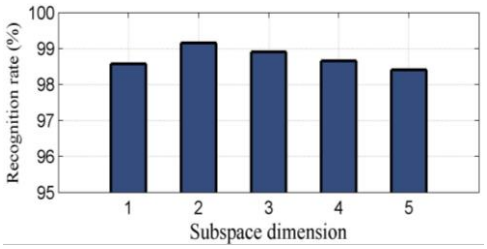


Fig. 4. Recognition performances versus different acoustic subspace dimensions

3.2.2. Validation of the proposed approach over full RWCP sound scene database

In the last experiment, our method achieved promising recognition performance on 12 categories of environmental sounds. In this part, we evaluated the proposed approach over full RWCP sound scene database with 105 sound categories and 9722 sample clips. We set the acoustic subspace dimension to 2 based on the last experiment. Finally, we achieved 94.41% recognition rate in 10-folder cross-validation. Besides, there are some duplicated categories in RWCP sound scene database, i.e. there are 4 kinds of phone rings and 5 types of bells. After merging the duplicated categories, we obtain 62 categories for classification, and the recognition rate reached to 96.67%.

4. CONCLUSIONS

An effective environmental sound recognition scheme is proposed in this study. At feature extraction stage, Gabor transform was employed to characterize the time-frequency distributions of environmental sounds. We further formulated the Gabor spectrogram with low rank subspace representation. The kernel Fisher discriminant analysis was applied over acoustic subspaces to characterize the class conditional distributions which are favorable for multi-class recognition. Experimental results over real world dataset validated the effectiveness of the proposed framework.

Acknowledgements: This work was partly supported by the Japan Science and Technology Agency (JST) A-STEP project.

12. REFERENCES

[1] Temko, A. and C. Nadeu, "Acoustic event detection in meeting-room environments", *Pattern Recognition Letters*, vol. 30, no.14, pp. 1281-1288, 2009

[2] Stavros Ntalampiras, Ilyas Potamitis, Nikos Fakotakis, "Exploiting Temporal Feature Integration for Generalized Sound Recognition", *EURASIP Journal on Advances in Signal Processing*, Article ID 594103, 2009

[3] Chu, S., Narayanan, S. and Kuo, C.C.J., "Environmental Sound Recognition With Time-Frequency Audio Features", *IEEE Trans. Audio, Speech, Lang Process.*, vol. 17, no.6, pp.1142-1158, 2009

[4] Courtenay V. Cotton and Daniel P. W. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection", *In Proc. of WASPAA 2011*, 2011, pp. 69-72

[5] Atrey, P. K., N. C. Maddage, et al., "Audio Based Event Detection for Multimedia Surveillance", *In Proc. of ICASSP2006*

[6] Cowling, M. and R. Sitte, "Comparison of techniques for environmental sound recognition", *Pattern Recognition Letters*, vol.24, no.15, pp. 2895-2907, 2003

[7] Guo, G., & Li, S. Z, "Content-based audio classification and retrieval by support vector machines", *IEEE Trans. on Neural Networks*, vol.14, no.1, pp. 209–215, 2003

[8] Yamakawa, N., Kitahara, T., Takahashi, T., Komatani, K., Ogata, T., Okuno, H.G., "Effects of modelling within-and between-frame temporal variations in power spectra on non-verbal sound recognition", *In Proc. of INTERSPEECH 2010*, 2010, pp. 2342–2345

[9] X. Zhuang, Z. Zhou, A. Hasegawa-J, and T.S. Huang, "Real-world acoustic event detection", *Pattern recognition Letters,* vol.31, no. 12, pp. 1543–1551, 2010.

[10] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, K.-R. Muller, "Fisher discriminant analysis with kernels", *In Proc. of Neural Networks for Signal Processing IX*, 1999, pp. 41–48.

[11] S. Li and A. Jain, *Handbook of Face Recognition*, Springer-Verlag, 2005.

[12] Takumi Kobayashi, "Generalized Mutual Subspace Based Methods", *In Proc of ACCV*, 2012

[13] A. Bjoerck and G.H. Golub, "Numerical methods for computing angles between linear subspaces", *Mathematics of Computation*, vol.27, no.123, pp. 579-594, 1973

[14] Real World Computing Partnership, "RWCP Sound Scene Database", http://tosa.mri.co.jp/sounddb/index.htm