AUDITORY CONTEXT CLASSIFICATION USING RANDOM FORESTS

Li Yang and Feng Su

State Key Laboratory for Novel Software Technology
Nanjing University
Nanjing, 210093, China

ABSTRACT

High-level semantic information can be extracted from audio materials to facilitate various content-based analysis and context-awareness applications. In this paper, we propose a novel automatic auditory context classification method, which combines the characterization of audio events and the inference of auditory context category in a single ensemble analysis framework. In the proposed framework, key audio events in the context are characterized by composite features from discriminative representation models (local discriminant bases, pseudo-semantic and bag-of-audio-words) learned from samples. A random forest based ensemble learning and classification model is employed for auditory contexts, in which individual segments of audio stream are classified and aggregated by Hough voting or bagging to form the final context category. The effectiveness of the proposed approach is demonstrated by the experimental results.

***Index Terms*—** Auditory context, random forest, local discriminant bases, MFCC, HMM

1. INTRODUCTION

Compared to visual signals, audio signals could be obtained and used in challenging conditions such as poor lighting or visual obstruction and are relatively more inexpensive in storing and computing, making it a common type of data in many exciting applications. Among others, as the widely spreading usage and the increasing performance for audio acquisition and computation, audio has played an important role in enhancing the context awareness in various systems, which imitates human's ability of understanding the high-level semantics of the auditory context. For example, given an audio stream comprising talks, laughter, music, clashing of knives, forks and dishes, we want a context-aware system could automatically infer that it would be more probable in a restaurant, rather than in a moving vehicle, just like an image of the same site can give us visually.

Auditory contexts here refer to the acoustic modeling of a specific location or site such as restaurant or bus station, while audio events are short audio segments with distinct acoustic patterns corresponding to specific object or event in an auditory context, such as laughing or gunfire. One major difficulty in auditory context inference, however, lies in the modeling of large amounts of environmental sounds constituting various audio events or background sounds. Different from common audio data types like speech or music, which has a formantic or harmonic spectral structure, environmental sounds are usually unstructured with a broad noise-like flat spectrum and diverse variety of signal composition. Moreover, the loose and possibly ambiguous connections between audio events and contexts require sophisticated modeling techniques.

Compared to general audio signal analysis that have received long time interests in the past years, research on unstructured environmental sound and auditory context are relatively less. Eronen et al. [1] developed a system to evaluate the recognition accuracy of various audio features (ZCR, Band-energy, Spectral flux, etc.), feature transforms (PCA, ICA, LDA), classifiers (k-NN and HMM) on 24 audio contexts, and compared the results with human listeners. Chu et al. [2] proposed to use the matching pursuit (MP) algorithm to select effective time-frequency features as the supplementation to Mel-frequency cepstral coefficients (MFCC) for audio environment characterization, in which general acoustic environment types are characterized as a whole rather than collections of discrete audio events pre-extracted. Cai et al. [3] modeled a set of key audio effects and several background sounds with HMMs, and proposed a Bayesian network-based approach to discover the high-level semantics of an auditory context embedded in key effect sequences. In [4], a hybrid SVM/KNN classifier is used for environmental sound classification based on MPEG-7 low-level audio descriptors.

In this paper, we propose an ensemble classification scheme for unstructured auditory contexts, which comprises two key ingredients: 1) the learning of discriminative feature representations for various audio events, and 2) the learning of the mapping between features of audio segments and the context using a random forest based ensemble classifier framework. The block diagram of the proposed algorithm is outlined in Fig. 1.

The remainder of the paper is organized as follows. Section 2 describes the audio feature extraction. Section 3 describes the random forest based classification model for au-

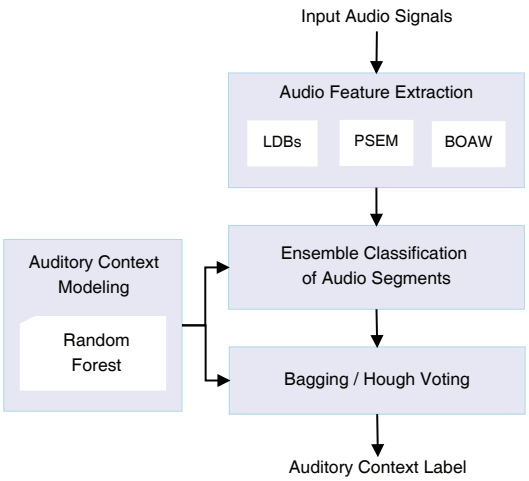

Fig. 1. Block diagram of the proposed auditory context classification algorithm.

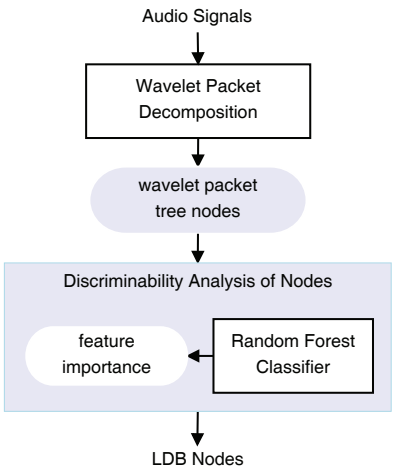

Fig. 2. Block diagram of LDB algorithm using random forest.

ditory contexts. In Section 4, we present the experimental results of the proposed method and some discussions.

2. FEATURE EXTRACTION

We employ a composite audio feature representation that characterizes different aspects of the signal stream at audio event scale. Three elementary features: the *local discriminant bases (LDB)* feature, the *pseudo-semantic (PSEM)* feature, and the *bag-of-audio-words (BOAW)* feature, are concatenated to form the feature vector for an audio signal segment. All elementary features rely on representation models learned from training samples of different audio event types.

2.1. Local Discriminant Bases (LDB) Feature

The LDB algorithm proposed by Saito and Coifman [5] is one kind of "best-basis" method to select an orthonormal optimum basis or subspace from a large collection of bases, which minimizes entropy or maximizes a certain discriminant measure among classes. Once such a basis is selected, a small number of most significant coordinates (features) can be used to enhance the performance of the classifier without losing important information of the problem.

In this work, we propose a random forest based expansion to the original LDB algorithm introduced by [6] for audio feature generation. The block diagram of our LDB algorithm is shown in Fig. 2, in which LDB is applied on the wavelet packet bases of the audio signals to identify the subspaces (i.e. subset of nodes) that different classes show high disparity.

Briefly, in LDB, an audio stream sample x_i of 2^u length is decomposed into a binary wavelet packets tree: $x_i = \sum_{j,k,l}[a_{j,k,l}]_i \cdot \mathbf{w}_{j,k,l}$, where j denotes the level of the tree, k denotes the node index in level j, $\{\mathbf{w}_{j,k,l}\}_{l=0}^{l=2^{u-j}-1}$ are the set of wavelet packet basis vectors and $a_{j,k,l}$ are the basis vector coefficients at node (j,k). To identify the best node subset providing maximum discriminative information between P classes, [6] computes the average dissimilarity values $D_{(j,k)}$ of node (j,k) on sample subsets of all $\binom{P}{2}$ combination of two different signal classes, and choose Q_{LDB} nodes with consistently (most frequently) high values of $D_{(j,k)}$ over multiple random subsampling trials as the selected LDB nodes. $D_{(j,k)}$ can take any form of dissimilarity measurements on values computed from the wavelet basis coefficients.

Since the essence of LDB is to evaluate the discriminability of each node, besides the standard dissimilarity-based LDB algorithm, we can replace the pair-wise assessment of dissimilarity on individual node in [6] with the holistic variable importance measurements computed by a dedicated random forest classifier [7] for feature selection, which is trained on the multi-class labeled audio event samples, as shown in Fig. 2. Specifically, for each training sample, we decompose it into a full J-level binary wavelet packets tree, catenate all the normalized energy values $E_{(j,k)}$ computed at every node (j,k) to form the raw feature vector, and use it with the sample's class label to train the random forest classifier. After training, the first Q_{LDB} variables (nodes) with maximum importance are chosen as the final LDB nodes. Then, for an input audio segment, we extract the feature values at these nodes to form the Q_{LDB}-dim feature vector.

The intrinsic multi-scale property and the adaptive learning of feature subspace in LDB are particularly useful in describing discriminative characteristics between audio events, which vary significantly in duration and spectral composition.

2.2. Pseudo-Semantic (PSEM) Feature

The pseudo-semantic feature is derived from a set of audio-event-specific HMM models, representing the intermediate characteristics between low-level physical audio features and

high-level semantic audio events. Specifically, for each audio event category, we extract sequences of MFCC features from its training samples, and use them to train one HMM model M_i ($i \in [1, C]$, C is the number of event categories) that captures the timbre and rhythm characteristics of the audio event.

For an input audio segment, the extracted MFCC sequence is given to every HMM in $\{M_i\}$, and the set of output likelihood values $\{l_1, l_2, \ldots, l_C\}$ by the Forward algorithm forms the C-dim PSEM feature vector.

2.3. Bag-of-Audio-Words (BOAW) Feature

The BOAW feature encodes the distribution of audio codewords of an input audio segment against one cross-class dictionary, in which the codeword is representative MFCC coefficients learned from samples. To build the dictionary, an unsupervised k-means algorithm is applied on the MFCC features extracted from all samples of all audio event categories and after convergence the final cluster number Q_{BOAW} is taken as the dimension of BOAW feature. The Q_{BOAW} clusters constitute the codewords in the dictionary.

For an input audio segment, we initialize a Q_{BOAW}-bins histogram $H_B(j)$ and extract a set of MFCC features $\{F_i\}$ from the signals. Next, each F_i is matched against the codeword dictionary, and denote its distance to j-th cluster by d_i^j ($j \in [1..Q_{BOAW}]$), then the j-th bin of H_B is increased by $\Delta_j = 1 - d_i^j / \sum_{n=1}^{Q_{BOAW}} d_i^n$. After all $\{F_i\}$ are processed, the normalized histogram is used as the BOAW feature.

3. AUDITORY CONTEXT CLASSIFICATION

In this section, we propose a random forest based method for auditory context classification. The main difference of the proposed method from other existing ones, which usually adopt a HMM-based model, is that we employ a *bag-of-event* model for auditory contexts, based on the observation that auditory contexts typically lack of obvious temporal evolution characteristics of the audio events, therefore we obviate explicit modeling of temporal correlation of events at the context level. Correspondingly, we propose an ensemble bagging/voting scheme for auditory context classification, which aggregates the class votes casted by individual segments of the input audio stream for the potential context category.

3.1. The Random Forest Model

Random forest is a collection of decision tree classifiers, each taking the input feature vector, classifying it and outputting its own vote on the possible class label. The final classification output of the whole forest is then the majority or other integration measurements of individual votes. Compared to a single decision tree or other simple classifier, random forest assembles together several trees trained in a randomized way and usually achieves superior generalization and stability.

In our work, a random forest model dedicated for context classification is used to learn the mapping between features of individual audio segments and their probabilistic votes on potential categories of the auditory context. Given samples of different auditory contexts, the training process is as follows:

1. Initialize a set of L decision trees $\{T_1, T_2, \ldots, T_L\}$ with required parameters (maximum depth, minimum sample count per node, etc.).
2. Divide each training audio sample into sequence of segments with sufficient overlapping, and extract the composite feature vector $\{\vec{v}_i\}$ for each segment.
3. Associate each feature vector $\vec{v}_i$ with the corresponding auditory context class label r_i ($r_i \in [1..K]$, K is the number of context classes) of the training sample.
4. Join all feature vectors together to form the data matrix $[\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_N]$, and use it with class labels $[r_1, r_2, \ldots, r_N]$ to train the random forest.

Once trained, the leaf nodes of forest trees encode the characteristic features of audio segment for specific context classes.

3.2. Classification By Random Forest

Given an input audio stream with unknown context label, the classfication using the random forest is as follows:

1. Decompose the stream into sequence of audio segments and extract features, as in the training process.
2. Send each segment down every tree in forest and collect the outputs at the predicted leaf node of every tree.
3. Aggregate the outputs to infer the potential auditory context class for the input stream.

For the aggregation of the forest outputs, we can use the *Hough voting* method that accumulates probabilistic votes casted by training samples. Let $R_{t,l} = \{s_i | i \in [1..N_{t,l}]\}$ is the set of training samples stored in the l-th leaf node of the t-th tree reached by an audio segment x_j. We initialize a 1D Hough vector $H_C(k)$ ($k = 1..K$), each element containing the accumulated votes for the corresponding context class. Then, for each sample s_i in $R_{t,l}$ whose class is k, we add a contribution to the bin $H_C(k)$. After all the input audio segments $\{x_j\}$ are processed, the peak in $H_C(k)$ indicates the MAP estimate for the potential context class. For higher efficiency, an alternative and simpler *bagging* method can be employed, which takes the majority class predicted by trees as the potential context category (i.e. the target bin in $H_C(k)$) for each input audio segment.

4. EXPERIMENTAL RESULTS

To evaluate the proposed auditory context classification method, we collect test data for 10 auditory contexts from Internet and some movie/TV clips, including 6 outdoor and 4 indoor contexts: auditorium, war field, forest, beach, train

Table 1. Confusion matrix for 10 auditory contexts.

Actual Classes	Classified Contexts % (in same order as rows)									
	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
1. audito.	72			2	8			18		
2. war		79			15	1				5
3. forest			100							
4. beach			2	62		1	3	19	1	12
5. station					94	1		3	2	
6. street					23	49	14	14		
7. vehicle					14		66	18		2
8. playgnd			1	7				89	3	
9. restau.	1							16	83	
10. rain			1							99

station, street, inside vehicle, playground, restaurant and raining. Total 21 audio event categories are considered, including engine, car-braking, siren, horn, gun-shot, explosion, dishware/cutlery, music, running water, bird twittering, thundering, talking, applause, laughter, cheer, etc. The training and testing set contain around 100 and 150 mono channel audio samples, respectively, for each audio event and context category. The typical sample length is 1-3s for audio events and 15s-2min for contexts with 44.1kHz sampling rate.

In experiment, we use a 1s analysis window with 50% overlapping on the audio stream for generating signal segments and extracting features. We learn a 20-dim LDB, 14-dim BOAW and 21-dim PSEM representation from audio event samples. MFCCs are extracted from every 1024 data points in the window. Each of the two random forest models, one trained with audio event samples for LDB feature selection and another trained with context samples for context classification, has 30 decision trees with max depth set to 15. Table 1 gives the confusion matrix of a typical trial of the proposed method with the composite features on the test set. Table 2 compares the average accuracy of the method with different feature compositions.

The experiment results show the effectiveness of the proposed random forest based framework, and combination of several heterogeneous features incrementally enhances the average performance. Meanwhile, as environmental auditory contexts are typically unstructured, the proposed discriminative parts-based model yields averagely higher (or comparable, for some contexts) performance in experiments than the more complicated temporal HMM context model. The flexibility in tuning feature compositions according to the complexity and efficiency requirements makes the framework potentially applicable to a wide range of circumstances.

5. ACKNOWLEDGMENTS

Research supported by the National Science Foundation of China under Grant Nos. 61003113 and 61021062.

Table 2. Performance of auditory context classification by using different features with the proposed random forest based context model. 'L+P' means the combination of LDB and PSEM feature, 'L+P+B' further incorporates BOAW feature. The last column shows the results of using the MFCC features with the proposed random forest based context model (the first) and a HMM-based context model (the second).

Category	Average Accuracy (%)					
	LDB	PSEM	BOAW	L+P	L+P+B	MFCC
audito.	72	70	71	72	73	72 / 73
war	78	65	63	76	79	75 / 74
forest	100	99	100	100	100	99 / 96
beach	27	44	50	53	63	45 / 47
station	89	90	85	90	93	93 / 88
street	43	38	37	52	48	45 / 58
vehicle	64	94	78	68	66	92 / 53
playgnd	87	81	73	90	90	81 / 72
restau.	83	95	85	84	84	87 / 82
rain	98	99	98	99	99	98 / 85
Average	**74.3**	**77.7**	**74.2**	**78.6**	**79.7**	**78.5 / 72.9**

6. REFERENCES

[1] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE TASLP*, vol. 14, no. 1, pp. 321–329, January 2006.

[2] S. Chu, S. Narayanan, and C.-C. Jay Kuo, "Environmental sound recognition with time-frequency audio features," *IEEE TASLP*, vol. 17, no. 6, pp. 1142–1158, August 2009.

[3] R. Cai, L. Lu, A. Hanjalic, H. Zhang, and L. Cai, "A flexible framework for key audio effects detection and auditory context inference," *IEEE TASLP*, vol. 14, no. 3, pp. 1026–1039, May 2006.

[4] J. Wang, J. Wang, K. He, and C. Hsu, "Environmental sound classification using hybrid SVM/KNN classifier and mpeg-7 audio low-level descriptor," in *IJCNN 2006*, 2006, pp. 1731–1735.

[5] N. Saito and R. R. Coifman, "Local discriminant bases and their applications," *J. of Mathematical Imaging and Vision*, vol. 5, no. 4, pp. 337–358, 1995.

[6] K. Umapathy, S. Krishnan, and R. K. Rao, "Audio signal feature extraction and classification using local discriminant bases," *IEEE TASLP*, vol. 15, no. 4, pp. 1236–1246, May 2007.

[7] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.