UNSUPERVISED AUDITORY SCENE CATEGORIZATION VIA KEY AUDIO EFFECTS AND INFORMATION-THEORETIC CO-CLUSTERING

Rui Cai[†*], *Lie Lu*[‡], *and Lian-Hong Cai*[†]

[†]Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China
[‡]Microsoft Research Asia, Beijing, 100080, China

ABSTRACT

Automatic categorization of auditory scenes is very useful in various content-based multimedia applications, such as video indexing and context-aware computing. In this paper, an unsupervised approach is proposed to group auditory scenes with similar semantics. In our approach, auditory scenes are described with the key audio effects they contained. In order to exploit the relationships between different audio effects and provide more accurate similarity measure for auditory scene categorization, co-clustering is utilized to group the auditory scenes and key audio effects simultaneously. In addition, Bayesian Information Criterion (BIC) is used to automatically select the cluster numbers for both the key effects and the auditory scenes. Evaluation on 272 auditory scenes extracted from 12-hour audio data shows very encouraging results.

1. INTRODUCTION

An auditory scene is a semantically consistent sound segment that is characterized by a few dominant sources of sound [2]. Automatic grouping auditory scenes with similar semantics is very useful for many multimedia applications, such as semantic event detection or indexing in videos [6][10] and context-aware computing [9].

To classify auditory scenes, some previous works establish a direct mapping from low-level audio features to high-level semantics. For example, in [9], *k*-NN classifier and GMM are built to classify auditory scenes into 26 pre-defined semantic categories, based on low-level features such as short-time energy, zero-crossing rate, LPC, and MFCC. However, those low-level features may vary significantly among various audio samples belonging to the same semantic category, and thus may lead to unsatisfying performance in practice.

To bridge the gap between low-level features and high-level semantics, *key audio effects* have been utilized as middle-level representations in the auditory scene categorization [6][10]. Key audio effects are those special effects playing critical roles in human's understanding of the auditory scene. In general, among those auditory scenes with similar semantics, there always exist some similar key audio effects. For instance, *cheer* and *laughter* are usually associated with *humor* scenes in comedies, and *explosion* and *gun-shot* often indicate *violence* scenes in action movies. Thus, based on the key audio effects, each auditory scene could be classified into one pre-defined class, either by heuristic rules [6] or by supervised statistical learning [10], assuming the semantic categories are known in *a priori*.

* This wok was performed when the first author was a visiting student in Microsoft Research Asia.

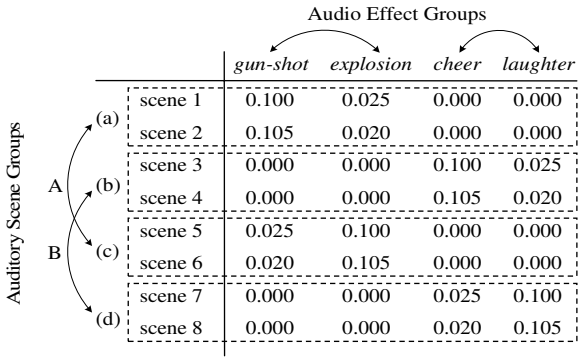

Fig. 1. Key audio effects based auditory scene categorization

However, in most cases, it is difficult to define all the semantic categories beforehand. Thus, unsupervised auditory scene categorization with key audio effects is highly expected. Traditional one-way clustering algorithm, such as *K*-means, usually can not work well in this problem. Fig. 1 illustrates such an example. In Fig. 1, there are 8 auditory scenes and each auditory scene is described by the occurrence probabilities of four key audio effects. With one-way clustering, all these key audio effects are considered independently in the scene's similarity measure, and thus there are four categories, as (a)-(d) show. But in fact, there are actually only two scene groups: (a) and (c) can be grouped to *A* which denotes the scenes of *war*, and (b) and (d) can be grouped to *B* which represents the scenes of *humor* in comedies. This is because *gun-shot* usually happens with *explosion* in *war* scenes; and *cheer* and *laughter* occur together in *humor* scenes, whatever their ratios are in the scenes. It indicates that in real world there are *group* phenomena among audio effects. That is, some audio effects usually occur together, while some others seldom happen subsequently. Those audio effects in the same effect group usually describe similar auditory scenes. Therefore in the scene grouping, the "distance" between two key effects in the same audio effect group should be smaller than that among different effect groups.

Thus to provide more reasonable results of semantic scenes clustering, key audio effects also need to be grouped according to their co-occurrences in auditory scenes. The clustering processes of auditory scenes and key audio effects are essentially dependent on each other. That is, the scene groups are relevant to the audio effect groups and vice versa. Similar cases appear in the document and keyword clustering for information retrieval. Here, auditory scenes can be taken as documents, and key audio effects are keywords. One solution proposed recently to solve the duality between the document and keyword clustering is the co-clustering algorithm [3][4], which clusters the document and keyword simultaneously. Two co-clustering approaches have

been proposed in literatures. One is based on the spectral graph partition [3] and the other utilizes information theory [4]. In our approach, the Information-Theoretic Co-Clustering is adopted to co-cluster the auditory scenes and key audio effects, since it has less restriction than the other one in practice.

Moreover, as the cluster numbers in current Information-Theoretic Co-Clustering algorithm are assumed to be known beforehand, in this paper, we extend the algorithm with Bayesian Information Criterion (BIC) [8] to automatically select the cluster numbers of both auditory scenes and key effects in clustering.

The rest of this paper is organized as follows. In Section 2, we present the process of the Information-Theoretic Co-Clustering based auditory scene categorization, as well as the selection of cluster numbers. In Section 3, experiments and evaluations are presented. The conclusion of our work is given in Section 4.

2. AUDITORY SCENE CATEGORIZATION

In this section, we formulate the problem of the audio key effect based auditory scene categorization under the scheme of co-clustering, and present the details of the Information-Theoretic Co-Clustering process. Then, a Bayesian Information Criterion based approach is proposed to automatically select the cluster numbers in the co-clustering.

2.1. Information-Theoretic Co-clustering

Suppose there are m auditory segments to be categorized, and n key audio effects are used to describe these scenes. All the auditory scenes could be considered as being generated by a discrete random variable S, whose value s is taken in the set $\{s_1, \ldots, s_m\}$. And similarly, all the key audio effects could be taken as being dominated by another discrete random variable E, whose value e is taken in the set $\{e_1, \ldots, e_n\}$. Let $p(S, E)$ denote the joint probability distribution between S and E. As S and E are both discrete, $p(S, E)$ is in nature an $m \times n$ matrix, whose element is represented as $p(s, e)$. Such a matrix is often called a two-dimensional *contingency table* or *co-occurrence table*. In $p(S, E)$, each row represents one auditory scene and each column denotes one key audio effect.

Suppose S and E could be grouped into k and l disjoint clusters, denoting as $\{s^*_1, \ldots, s^*_k\}$ and $\{e^*_1, \ldots, e^*_l\}$ respectively. These clusters could also be regarded as being generated by two discrete random variables S^* and E^*.

In the view of information theory, a fundamental quantity that measures the amount of information shared between S and E is the *mutual information* $I(S; E)$.

$$I(S;E) = \sum_s \sum_e p(s,e) \log_2 \frac{p(s,e)}{p(s)p(e)} \tag{1}$$

In [4], it indicates that an optimal co-clustering should minimize the *loss of mutual information* after clustering, *i.e.* the optimal clusters should satisfy

$$\arg\min_{(S^*,E^*)} (I(S;E) - I(S^*;E^*)) \tag{2}$$

The *loss of mutual information* can be represented as

$$I(S;E) - I(S^*;E^*) = KL(p(S,E), q(S,E)) \tag{3}$$

where $q(S,E)$ is also a distribution in the form of an $m \times n$ matrix

$$q(s,e) = p(s^*,e^*)\, p(s \mid s^*)\, p(e \mid e^*), \text{ where } s \in s^*, e \in e^* \tag{4}$$

and $KL(f, g)$ denotes the *Kullback-Leibler* (*K-L*) *divergence* or *relative entropy* of two distributions $f(x)$ and $g(x)$, as

$$KL(f,g) = \sum_x f(x) \log_2 \frac{f(x)}{g(x)} \tag{5}$$

To facilitate the clustering process, Eq. (3) can be further expressed as Eq. (6) and (7) in a symmetrical manner [4].

$$KL(p(S,E), q(S,E)) = \sum_s \sum_{s \in s^*} p(s) KL(p(E \mid s), q(E \mid s^*)) \tag{6}$$

$$KL(p(S,E), q(S,E)) = \sum_e \sum_{e \in e^*} p(e) KL(p(S \mid e), q(S \mid e^*)) \tag{7}$$

From Eq. (6) and (7), it shows that the minimizing of the *loss of mutual information* can be achieved by minimizing the *K-L* divergence between $p(E|s)$ and $q(E|s^*)$, as well as the divergence between $p(S|e)$ and $q(S|e^*)$. Thus an iterative co-clustering algorithm could be carried out as following four steps:

1) Initialization: Assigning all the auditory scenes into k parts, and the key effects into l parts. Then calculate the initial value of the q matrix.
2) Update row clusters: First, for each row s, find its new cluster index i in the measure of *K-L* divergence, as

   $$i = \arg\min_k KL(p(E \mid s), q(E \mid s^*_k)) \tag{8}$$

   Thus the *K-L* divergence of $p(E|s)$ and $q(E|s^*)$ is decreased in this step. With the new cluster indices of rows, update the q matrix according to Eq. (4).
3) Update column clusters: Based on the updated q matrix in step 2, find a new cluster index j for each column e in the measure of *K-L* divergence, as

   $$j = \arg\min_l KL(p(S \mid e), q(S \mid e^*_l)) \tag{9}$$

   Thus the *K-L* divergence of $p(S|e)$ and $q(S|e^*)$ is decreased in this step. With the new cluster indices of columns, update the q matrix again.
4) Re-calculate the *loss of mutual information* by Eq. (3). If the change in the *loss of mutual information* is smaller than a pre-defined threshold, stop the iteration process and return the clustering results; otherwise go to step 2 to start a new iteration.

In [4], it has been proved that the above iteration process could monotonically decrease the *loss of mutual information* and guarantee to converge to a local minimum. In implementation, the maximally far apart criterion is used to select the initial cluster centers, and the local search strategy is utilized to increase the quality of the local optimal [5]. The algorithm is computationally efficient and its complexity is $O(n \cdot \tau \cdot (k+l))$, where n is the number of nonzeros in $p(S, E)$ and τ is the iteration number.

2.2. Estimation of the Cluster Numbers

The row cluster numbers k and the columns cluster number l are assumed to be known in the above co-clustering algorithm. However, in most applications, it's hard to precisely specify the cluster numbers beforehand.

As mentioned in Section 2.1, the criterion used to evaluate the clustering results is the *loss of mutual information*. However, according to the definition, the *loss of mutual information* has its inherent variation trend with the change of cluster numbers, as illustrated in Fig. 2. In Fig. 2, it is noted that more clusters are used, more *mutual information* is reserved. For example, when

both the row and column cluster numbers are one, 100% *mutual information* loses after the clustering; while when the cluster number is equal to the original sample amount, no *mutual information* loses. Although the *loss of mutual information* decreases with more clusters, the model complexity (the number of parameters in the model) increases significantly.

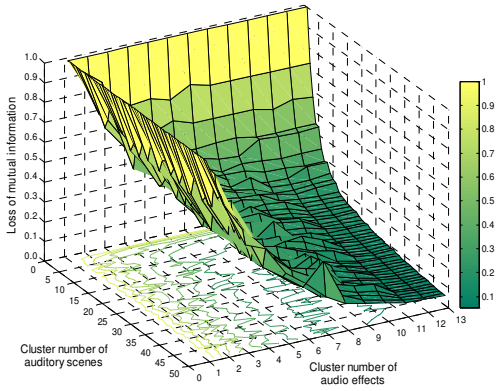

Fig. 2. Illustration of the *loss of mutual information* with different cluster numbers of auditory scenes and key audio effects

In this paper, to balance the *loss of mutual information* and the model complexity, we utilize the Bayesian Information Criterion (BIC) [8] to select the optimal cluster numbers of co-clustering. BIC has been successfully used to automatically select the cluster number for *K*-means clustering [1]. Given a model, BIC trades off the data likelihood *L* with the model complexity $|\Theta|$. In practice, the former has a weighting factor λ; and the latter is modulated by the logarithm of the total number of samples *T* in the database, as

$$BIC = \lambda L - \frac{1}{2} |\Theta| \log(T) \tag{10}$$

In our scheme of co-clustering, given values of *k* and *l*, the data likelihood *L* in Eq. (10) could be described by the logarithm of the ratio between the *mutual information* after clustering ($I(S^*; E^*)$) and the original *mutual information* ($I(S; E)$). It is assumed that the model reserving more *mutual information* would have higher "probability" to fit the data. Meanwhile, as co-clustering is a two-way clustering, the model complexity here should consist two parts: the size of the row clusters (*k* cluster centers of *n* dimensionality) and the size of the column clusters (*l* cluster centers of *m* dimensionality). Thus the definition of the BIC in our algorithm is designed as Eq. (11).

$$BIC(k,l) = \lambda \log \frac{I(S^*;E^*)}{I(S;E)} - (\frac{nk}{2} \log m + \frac{ml}{2} \log n) \tag{11}$$

In implementation, λ is set experimentally as $m \times n$. The algorithm searches over all the (k, l) pairs in a pre-defined range, and the model with the highest BIC score is chosen as the optimal clustering result.

3. EVALUATIONS

The evaluations of the proposed algorithm have been performed on 272 auditory scene segments extracted from about 12-hour audio tracks, including movies and entertainment TV shows. All the audio streams are in 16 KHz, 16-bit and mono channel.

In these auditory scenes, we detect ten key audio effects including *applause*, *car-racing*, *cheer*, *car-crash*, *explosion*, *gun-shot*, *helicopter*, *laughter*, *plane*, and *siren*, and three general audio effects as *music*, *speech*, and *noise*. To evaluate the unsupervised clustering results, all the 272 auditory scenes are manually labeled into five semantic categories, including *excitement*, *humor*, *pursuit*, *fight* and *air-attack*. The key audio effects which possibly occur in each scene category are listed in Table 1.

Table 1. Relationships between the key audio effects and the auditory scene categories in the database.

Semantic Category	Key Audio Effects
excitement	*cheer*, *applause*
humor	*laughter*, *applause*
pursuit	*car-crash*, *car-racing*, *siren*, *helicopter*, *gun-shot*, *explosion*
fight	*gun-shot*, *explosion*
air-attack	*plane*, *explosion*

For each auditory scene, the framework proposed in our previous work [7] is utilized to detect the audio effects it contains. In the framework, Hidden Markov Models (HMMs) are used to model the audio effects. Then a sliding window of 1 second moves through the input audio stream with 0.5 second overlapping, and each window is further compared against each HMM and corresponding log-likelihood score (confidence) is obtained. At last, each window is classified into the audio class with the highest score.

Based on the detection results, we estimate the occurrence probability of the j^{th} key effect e_j in the i^{th} auditory scene, as:

$$P_{occur}(i,j) = \frac{1}{T_i} \sum_{1 \le t \le T_i, t \in e_j} c_j(t) \tag{12}$$

where T_i is the number of sliding windows in the i^{th} auditory scene, and $c_j(t)$ is the confidence score of the t^{th} sliding window belonging to audio effect e_j. Then each element $p(s, e)$ of the *contingency table p(S, E)* can be calculated as:

$$p(s_i, e_j) = \frac{P_{occur}(i,j)}{\sum_{1 \le i \le m} \sum_{1 \le j \le n} P_{occur}(i,j)} \tag{13}$$

To illustrate the efficiency of exploiting the relationships among various audio effects in the unsupervised auditory scene grouping, we compare the proposed co-clustering algorithm with those traditional one-way clustering algorithms. Here, the X-means algorithm [1], in which BIC is used to estimate the cluster number of *K*-means, is adopted in comparison. In the X-means clustering, we search the proper cluster number *K* in the range of ($1 \le K \le 50$); and in the Information-Theoretic Co-Clustering, we look for the auditory scene cluster number *k* and the audio effect cluster number *l* in the range of ($1 \le k \le 50$, $1 \le l \le 13$).

In experiments, we finally get 8 auditory scene categories by using the Information-Theoretic Co-Clustering, and 13 scene categories by using the X-means clustering. The detailed clustering results of the two algorithms are listed in Table 2 and Table 3 respectively. In Table 2 and Table 3, each row represents one obtained cluster and corresponding samples from each semantic category in the ground truth. To give a more explicit illustration of the performances, we manually group those clusters associated to the same ground truth category (as the shadow parts illustrated in Table 2 and Table 3), and then calculate corresponding precisions and recalls.

By comparing Table 2 and Table 3, it's clear that in general the co-clustering algorithm can achieve better performance in the auditory scene categorization. First, the number of auditory cate-

gories obtained by co-clustering is more close to the ground truth than that achieved with the X-means clustering. It indicates co-clustering can give a more exact approximation to the actual groups existing among the auditory scenes in database. Second, for most auditory categories, co-clustering can get higher precisions and recalls than the X-means algorithm. Averagely around 88.6% auditory scenes are correctly classified with the co-clustering algorithm, and the performance of the X-means clustering is just 83.82%.

Table 2. The clustering results obtained using the Information-Theoretic Co-Clustering algorithm

No.	*excitement*	*humor*	*pursuit*	*fight*	*air-attack*	**precision**
1	18	0	1	0	0	96.43%
2	9	0	0	0	0	
3	3	13	0	1	0	88.37%
4	1	25	0	0	0	
5	0	0	41	2	0	93.33%
6	0	0	29	3	0	
7	0	0	9	81	0	90.00%
8	0	0	9	2	25	69.44%
recall	87.10%	100.00%	78.65%	91.01%	100.00%	

Table 3. The clustering results obtained using the X-means clustering algorithm

No.	*excitement*	*humor*	*pursuit*	*fight*	*air-attack*	**precision**
1	21	0	0	0	0	100.00%
2	10	38	2	0	0	76.00%
3	0	0	22	0	0	92.96%
4	0	0	11	1	0	
5	0	0	24	3	0	
6	0	0	9	1	0	
7	0	0	1	16	1	83.51%
8	0	0	0	36	0	
9	0	0	3	9	0	
10	0	0	2	2	2	
11	0	0	7	18	0	
12	0	0	1	0	15	66.67%
13	0	0	7	3	7	
recall	67.74%	100.00%	74.16%	91.01%	88.00%	

Investigating in more detail, the clusters obtained with the Information-Theoretic Co-Clustering are more concentrated and consistent corresponding to the ground truth category. In contrast, samples in some clusters obtained with the X-means clustering are divergence. For example, the 2nd cluster in Table 3 consists of 10 samples from *excitement*, 38 samples from *humor*, and 2 samples from *pursuit*. This indicates that co-clustering can well characterize the difference among various scene groups by exploiting the relations among key audio effects.

Furthermore, with the co-clustering algorithm we also obtain seven key audio effect groups in the experiments, as shown in Table 4. These clustering results are basically consistent with human knowledge and our assumptions in Table 1. For example, by investigating the database, we found that the effects of *helicopter* and *siren* usually occur together with tense *music* in some *pursuit* scenes, and they are correctly grouped together with the co-clustering algorithm.

Table 4. The key audio effect groups obtained using the Information-Theoretic Co-Clustering

No.	Key Audio Effects	No.	Key Audio Effects
1	*helicopter, music, siren*	5	*laughter, cheer*
2	*speech*	6	*applause*
3	*plane*	7	*noise*
4	*car-racing, car-crash, explosion, gun-shot*		

4. CONCLUSION

This paper presented an unsupervised solution to auditory scene categorization by using key audio effects and the Information-Theoretic Co-Clustering. Co-clustering could provide a more reasonable similarity measure in auditory scene grouping by exploiting the relationships among various key audio effects. In addition, to automatically select the cluster numbers in the clustering, a Bayesian Information Criterion-based strategy is also proposed in this paper. Experiments show that the two-way Information-Theoretic Co-Clustering algorithm can achieve better performance than the traditional one-way *K*-means algorithm in the key audio effect based auditory scene categorization.

5. REFERENCES

[1] D. Pelleg and A.W. Moore, "X-means: Extending *K*-means with Efficient Estimation of the Number of Clusters," *Proc. of the 17th International Conference on Machine Learning*, pp.727-734, Stanford, CA, USA, Jun. 29-Jul. 2, 2000.

[2] H. Sundaram and S.-F. Chang, "Audio Scene Segmentation Using Multiple Feature, Models and Time Scales," *Proc. of IEEE International Conference on Acoustic, Speech and Signal Processing*, Vol.4, pp.2441-2444, Istanbul, Turkey, Jun. 5-9, 2000.

[3] I.S. Dhillon, "Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning," *Proc. of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.269-274, San Francisco, CA, USA, Aug. 26-29, 2001.

[4] I.S. Dhillon, S. Mallela, and D.S. Modha, "Information-Theoretic Co-clustering," *Proc. of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.89-98, Washington, DC, USA, Aug. 24-27, 2003.

[5] I.S. Dhillon and Y. Guan, "Information Theoretic Clustering of Sparse Co-Occurrence Data," *Proc. of 2003 IEEE International Conference on Data Mining*, pp.517-520, Melbourne, FL, USA, Nov. 19-22, 2003.

[6] M. Xu, N. Maddage, C.-S. Xu, M. Kankanhalli, and Q. Tian, "Creating Audio Keywords for Event Detection in Soccer Video," *Proc. of IEEE International Conference on Multimedia and Expo*, Vol.2, pp.281-284, Baltimore, MD, USA, Jul. 6-9, 2003.

[7] R. Cai, L. Lu, H.-J. Zhang, and L.-H. Cai, "Highlight Sound Effects Detection in Audio Stream," *Proc. of IEEE International Conference on Multimedia and Expo*, Vol.3, pp.37-40, Baltimore, MD, USA, Jul. 6-9, 2003.

[8] R.E. Kass and L. Wasserman, "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion," *Journal of the American Statistical Association*, Vol.90, No.431, pp.928-934, 1995.

[9] V. Peltonen, J. Tuomi, A.P. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational Auditory Scene Recognition," *Proc. of IEEE International Conference on Acoustic, Speech and Signal Processing*, Vol.2, pp.1941-1944, Orlando, USA, May 13-17, 2002.

[10] W.-H. Cheng, W.-T. Chu, and J.-L. Wu, "Semantic Context Detection based on Hierarchical Audio Models," *Proc. of the International Workshop on Multimedia Information Retrieval*, pp.109-115, Berkeley, CA, USA, Nov. 7, 2003.