Co-clustering for Auditory Scene Categorization

Rui Cai, Lie Lu, *Member, IEEE*, and Alan Hanjalic, *Member, IEEE*

***Abstract*—Auditory scenes are temporal audio segments with coherent semantic content. Automatically classifying and grouping auditory scenes with similar semantics into categories is beneficial for many multimedia applications, such as semantic event detection and indexing. For such semantic categorization, auditory scenes are first characterized with either low-level acoustic features or some mid-level representations like *audio effects*, and then supervised classifiers or unsupervised clustering algorithms are employed to group scene segments into various semantic categories. In this paper, we focus on the problem of automatically categorizing audio scenes in unsupervised manner. To achieve more reasonable clustering results, we introduce the co-clustering scheme to exploit potential grouping trends among different dimensions of feature spaces (either low-level or mid-level feature spaces), and provide more accurate similarity measure for comparing auditory scenes. Moreover, we also extend the co-clustering scheme with a strategy based on the Bayesian information criterion (BIC) to automatically estimate the numbers of clusters. Evaluation performed on 272 auditory scenes extracted from 12-h audio data shows very encouraging categorization results. Co-clustering achieved a better performance compared to some traditional one-way clustering algorithms, both based on the low-level acoustic features and on the mid-level audio effect representations. Finally, we present our vision regarding the applicability of this approach on general multimedia data, and also show some preliminary results on content-based image clustering.**

***Index Terms*—Audio content analysis, auditory scene categorization, co-clustering, local feature grouping trends.**

I. Introduction

An auditory scene is a semantically consistent audio segment that describes a certain event scenario or provides some contextual information [1]. For example, in action movies, an audio segment containing the sounds of *siren*, *car-racing*, and *car-crash* construct an auditory scene which may indicate a *pursuit* event; while an audio segment containing the sounds of *noise*, *speech*, and *car-horn* can be seen as an indication of a *street* scene context.

Automatically classifying or grouping auditory scenes with similar semantics into categories is useful for many multimedia applications. Many recent research works explored the role of auditory scenes for semantic event detection or indexing in audiovisual data (video) [2]–[6]. For example, audio scene analysis has been employed to detect scenes like *fighting*, *pursuit*, and *horror* in movies [2], [3], [5], [6], and highlight events like *goal* and *foul* in sports videos [4]. Other newly emerging applications of auditory scene categorization are in context-aware computing and pervasive computing [7]–[12]. As audio signals are easy to be recorded and stored on diverse portable devices, these collected audio data can be analyzed to recognize users' location and application/situation context. For instance, attempts were made in [11], [12] to automatically create E-diaries through analyzing and annotating a whole day's audio recording. Also mobile phone manufacturers have recognized the potential of recognizing representative environmental scenarios by classifying audio inputs [7], [9].

The features typically used to describe auditory scenes include both low-level acoustic features and mid-level representations. Some previous works established a direct mapping from low-level features to high-level semantics. For example, in [7], [9], low-level features such as short-time energy (STE), zero-crossing rate (ZCR), linear predictive coding (LPC) coefficients, and Mel frequency cepstral coefficients (MFCC) are used to characterize scene samples; while in [11], the entropy of energy distribution of discrete Fourier transform (DFT) spectrum is employed to depict audio signals. These low-level features can provide temporal and spectral descriptions of audio data, and are reported to be powerful in distinguishing some related simple scenes like *indoor/outdoor* and *speech/music*. However, for scenes with more complex semantics, such low-level features usually vary significantly among the scene samples belonging to a same semantic category, and thus may lead to unsatisfying categorization performance. To compensate for this deficiency of low-level features, some mid-level representations or *audio effects* have been utilized to characterize complex auditory scenes. Audio effects can be defined as specific elementary audio events that provide critical clues for human understanding of an auditory scene [2], [3], [6]. Examples of such effects are *cheer* and *laughter* in *humor* scenes in comedies, and *explosion* and *gun-shot* in *violence* scenes in action movies. As audio effects contain some semantic meaning, they can provide more exact and useful information to infer the semantic categories of complex audio scenes.

To categorize auditory scenes, most current works utilize supervised classifiers or rules, applied either to low-level acoustic features or mid-level representations. For example, K-NN classifiers and Gaussian mixture models (GMM) are built in [7] to classify auditory scenes into 26 predefined semantic categories. In [4], heuristic rules such as "if *double whistling*, then *Foul* or *Offside* happens" are used to infer the semantics of scenes in soccer games. Also, statistical models like support vector machine (SVM) and Bayesian network (BN) [3] have been employed to learn the relationships between audio effects and scene semantics. The supervised classification methods typically applied in the abovementioned approaches have

Manuscript received October 30, 2006; revised January 9, 2008. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Arnold Smeulders.

R. Cai and L. Lu are with Microsoft Research Asia, Beijing 100080, China (e-mail: ruicai@microsoft.com; llu@microsoft.com).

A. Hanjalic is with the Department of Mediamatics, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: A.Hanjalic@ewi.tudelft.nl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2008.921739

1520-9210/$25.00 © 2008 IEEE

been proved to be powerful in capturing the available prior knowledge and modeling complex mappings between features and high-level semantics, and gave encouraging results. In supervised approaches, the semantic categories are assumed to be known in *a priori*. This assumption, however, is not always valid, as for instance in the applications like pervasive computing [11]. There, the semantic scenes are unknown in advance, and it is therefore impossible to collect training data and learn proper supervised models or rules. Moreover, the applicability of most supervised approaches is limited to narrow application domains, as their effectiveness relies heavily on the quality and completeness of training data [13]. Therefore, not surprisingly, there is an increasing need for novel unsupervised scene categorization approaches, as also suggested in [14]. The development of such an approach taking as input both the low-level acoustic features and mid-level audio effects, is the focus of this paper.

Clustering theory [15] provides the most intuitive framework for grouping similar scenes to a semantic category in an unsupervised manner. In traditional one-way clustering algorithms such as K-means, similarities among scenes are estimated by measuring the distances among their relevant points in feature space, assuming that each feature gives equal contribution to the distance measure. However, such assumption is not always satisfied in practice, and usually leads to suboptimal clustering performance. In this paper, we first investigate possible ***local grouping trends*** among various features given the data distribution in the feature space, and also explain how such trends will affect the similarity measure during clustering. To characterize such grouping trends among features and use them to provide more reasonable clustering results, we propose a co-clustering based scheme to categorize auditory scenes. Co-clustering (also referred to as "bi-clustering") is a two-way simultaneous clustering algorithm, which has already been employed in research fields like bioinformatics [16], [17] and text analysis [18]–[20], mostly acting as a tool for generating co-occurrence statistics. In this paper, we show that co-clustering can also provide a framework to formulate the local feature grouping relations, and can provide more reasonable scene categorization results than one-way clustering approaches. Moreover, as the cluster numbers in the existing co-clustering algorithms are assumed to be known beforehand, we also generalize the co-clustering scheme by applying the Bayesian information criterion (BIC) to automatically select the optimal cluster numbers. Experiments on auditory scene categorization show very promising results. The co-clustering scheme achieved a better performance on both classification accuracy and cluster number prediction. We also discuss the applicability of our approach on other mediums than audio, and examine the performance with a preliminary evaluation on image clustering.

The rest of this paper is organized as follows. In Section II, we first unravel the phenomenon of local grouping trend among features and analyze how such trends affect a similarity measure. In Section III, we present our approach to auditory scene categorization based on the co-clustering scheme, as well as the BIC-based estimation of the cluster numbers. The evaluation of the proposed approach is presented in Section IV, while in Section V we discuss the applicability of the proposed approach to general multimedia data. Finally, conclusions and recommendations can be found in Section VI.

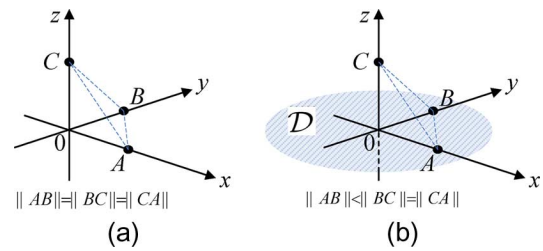

Fig. 1. Local feature grouping trend and its influence to the distance measure.

II. Analysis of Local Grouping Trends Among Features

It is known from the field of pattern classification that different features (or feature combinations) may provide different contributions in distinguishing samples belonging to different classes. Various algorithms such as the Bayesian structural EM [21], can be utilized to learn this information and estimate optimal model parameters for improving classification accuracy. However, in the unsupervised case, the absence of training samples makes it difficult to discover such latent relations between features and data clusters, although these relations exist and also affect the categorization performance. More specifically, for the problem addressed in this paper, given the auditory scenes and their representations in the feature space, a fundamental step in the scene categorization process is to measure the distance or similarity between any two scene samples. Without prior knowledge, each individual feature measured in a scene will contribute equally to the distance measure, what in practice may not always be a reasonable approach.

In this paper, we aim at improving the reliability of scene clustering results in a general case by exploring the relations between features and data in an unsupervised manner. In particular, we investigate how the relations between features change locally depending on various potential data clusters in the feature space. We will refer to such feature relations further as the "local feature grouping trends". In this section, we will first show some examples of such trends and their influence on distance/similarity computation, both in terms of low- and mid-level features, and then discuss why dealing with this phenomenon is difficult in the specific context of unsupervised auditory scene categorization.

A. Grouping Trends Between Low-Level Features

Fig. 1 illustrates an example of a local feature grouping trend and its influence on a distance measure. Fig. 1(a) shows three data points A, B, and C located on the $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ axes of a 3-dimensional feature space, respectively, and all at the same distance from the origin $\boldsymbol{o}$. If $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ are independent of each other, then we have $\|AB\| = \|BC\| = \|CA\|$. Now, we assume that it has been known that in the data set there is a cluster whose data points lay on the $\boldsymbol{x}-\boldsymbol{y}$ surface, as indicated by the region $\mathcal{D}$ in Fig. (1b), where A and B belong to cluster $\mathcal{D}$ while C does not. Then, in order to properly reveal the cluster $\mathcal{D}$ after clustering, the distance from the point C to each of the points A or B should be larger than the distance between A and B, i.e.,

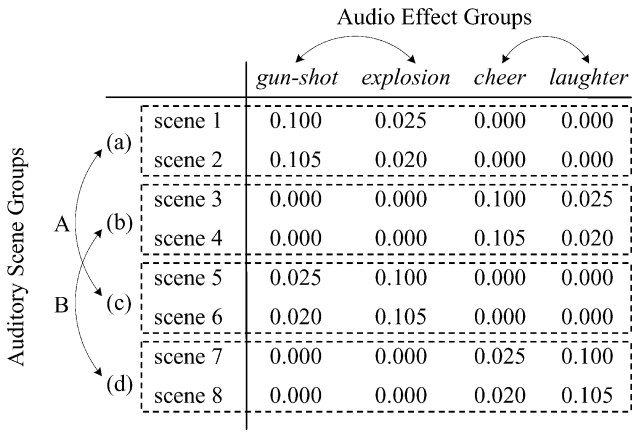

Fig. 2. Audio effects based auditory scene categorization.

$\|AB\| < \|BC\| = \|CA\|$, by means of the used distance measure. In other words, we can say that the dimensions $\boldsymbol{x}$ and $\boldsymbol{y}$ are apt to be ***locally grouped*** or correlated given the distribution of data cluster $\mathcal{D}$, and such local $\boldsymbol{x} - \boldsymbol{y}$ grouping should be considered in the distance measure to properly reveal the cluster $\mathcal{D}$. In the clustering practice, however, such local grouping trends among features cannot be analyzed in advance as data clusters like $\mathcal{D}$ are not known. As a matter of fact, revealing such clusters is the purpose of clustering.

B. Grouping Trends Between Audio Effects

The phenomenon similar to local feature grouping can also be found in the spaces representing mid-level audio effects. Understanding such grouping trends can even be more intuitive than for the abstract low-level feature case discussed above.

We consider the example in Fig. 2. This case involves 8 auditory scenes, each of which is described by the occurrence probabilities of four audio effects. With one-way clustering, all these four audio effects are considered independently in the scene's similarity measure, which results in four scene categories indicated by (a)–(d). However, a semantic analysis of the data suggests that the abovementioned grouping leads to over-segmentation of the content, and that grouping the scenes into only two categories would be more appropriate. The first of these categories, indicated as A, can be labeled as *war*, and consists of scene groups (a) and (c). The second one, the category B, represents *humor* and includes the scene groups (b) and (d). Here, we can say the dimensions of *cheer* and *laughter* are ***locally grouped*** given the samples from the *humor* scene. To discover higher-level semantic categories automatically, an algorithm should learn that a *gun-shot* usually occurs together with *explosion* in *war* scenes, while *cheer* and *laughter* often co-occur in *humor* scenes. More generally, there is the need in the development of categorization algorithms for effectively exploiting the fact that there are combinations of audio effects that often explain the semantics of audio data much better than the effects taken individually.

C. Discussions

The phenomena in Sections II-A and II-B can be generally summarized as the dependences between the processes of categorizing auditory scenes and discovering the local grouping trends among feature.[1] In other words:

- To achieve more reasonable semantic clusters with scene categorization, the local grouping trends among various features should be first exploited, to provide more accurate similarity measure to scene samples.
- Vice versa, to find out such local feature grouping trends existing in the data set of scene samples, the scene categories should be known beforehand to analyze the characteristics of local data distributions.

In supervised learning algorithms, such dependences have been embedded in models through analysis of the training data. However, in our unsupervised situation, this "chicken and egg" situation makes it difficult for the traditional clustering mechanisms, like K-means, to exploit such local grouping trends and optimize the results of auditory scene categorization.

As a side note, we wish to emphasize the difference between the discovery of local grouping trends among features and some well-known statistical data pre-processing mechanisms, such as principal component analysis (PCA) [22] and independent component analysis (ICA) [23]. PCA and ICA have an entirely different objective, namely searching for global correlations among features on the whole data set. In general, PCA and ICA are utilized to identify new meaningful underlying features and reduce the data dimensionality.

III. Co-clustering for Auditory Scene Categorization

As discussed in Section II, the local grouping trends, existing among features, will influence the clustering performance of auditory scenes. To achieve more reasonable clusters, such local grouping trends should be discovered and utilized in the distance measurement between scene samples.

As the local grouping trends depend on the features' co-occurrence in samples from various scene categories, an intuitive solution to reveal such trends is clustering features based on their co-occurrence first. The revealed relations are then used in measuring similarities among scene samples, to group scenes into different semantic categories. Based on the obtained scene groups, the group relations among features can be refined through a second-round clustering, and the refined feature groups can then again be employed to give more reasonable scene groups. However, for the above iterative clustering process the convergence is difficult to prove theoretically [24], and there is no plausible criterion, based on which such iterative process can be stopped.

To solve the abovementioned problems, in this paper, we propose a co-clustering based scheme. Co-clustering provides the possibility for a simultaneous two-way clustering, which was already proved to converge toward a local minimum [15], thus we can perform a co-clustering to reveal the local feature groups and the semantic scene categories, simultaneously. Two different co-clustering approaches have been proposed in recent literature. One is based on the spectral graph partition [18], and the other employs an information theoretic approach [19], [20]. In our scheme, the information-theoretic co-clustering (ITCC)

[1]In the following sections, when we refer to features we refer jointly to both low- and mid-level features (effects).

is adopted, since it imposes fewer restrictions than the other solution.[2]

In this section, we first formulate the problem of auditory scene categorization using the co-clustering scheme, and present the details of the information-theoretic co-clustering process. Then, a Bayesian information criterion based strategy is proposed to extend the ITCC scheme by providing the possibility to set the number of clusters automatically.

A. Information-Theoretic Co-clustering

Suppose there are m auditory scenes to be categorized, and n features are used to describe these scenes. The scenes can be considered as being generated by a discrete random variable S, whose value s is taken from the set $\{s_1, \cdots, s_m\}$. Similarly, the features can be assumed dominated by another discrete random variable F, whose value f is taken in from the set $\{f_1, \cdots, f_n\}$. Let $p(S, F)$ denote the joint probability distribution between S and F. As S and F are both discrete, $p(S, F)$ is an $m \times n$ matrix, whose elements can be represented by $p(s, f)$. Such a matrix is often called a 2-D contingency table or co-occurrence table [25]. In $p(S, F)$, each row represents one auditory scene and each column denotes one feature. We now also suppose that S and F could be grouped into k and l disjoint clusters, denoting as $\{s_1^*, \cdots, s_k^*\}$ and $\{f_1^*, \cdots, f_l^*\}$, respectively. These clusters could also be regarded as being generated by two discrete random variables S^* and F^*.

We start the information-theoretic approach to co-clustering by measuring the amount of information shared between S and F, that is, the mutual information $I(S; F)$, as

$$I(S;F) = \sum_s \sum_f p(s,f) \log_2 \frac{p(s,f)}{p(s)p(f)}. \tag{1}$$

The mutual information is taken as a measurement of the original information of the data collection, controlled by the latent relations between the variables S and F. The criterion of co-clustering is that such information should not be changed too much during the clustering, as the objective of the clustering is just to reveal such latent relations between the two variables. Based on such assumption, it was shown in [19] that an optimal co-clustering should minimize the loss of mutual information after clustering, i.e., for the optimal clusters we can write:

$$(\hat{S^*}; \hat{F^*}) = \arg \min_{S^*, F^*} I(S;F) - I(S^*;F^*). \tag{2}$$

The loss of mutual information can be represented as

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

[2]In spectral graph partition-based co-clustering, the numbers of clusters in both feature dimensions and samples should be same. Such assumption is too strict in the context of our work.

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

$$q(s,f) = p(s^*, f^*)p(s|s^*)p(f|f^*) \quad where \quad s \in s^*, f \in f^* \tag{4}$$

and where $KL(f, g)$ denotes the Kullback–Leibler (K–L) divergence or relative entropy of two distributions $f(x)$ and $g(x)$:

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

Also as shown in [19], the K-L divergence in (3) can be modified into (6) or (7):

$$KL(f,g) = \sum_{s^*} \sum_{s \in s^*} p(s) KL\left(p(F|s), q(F|s^*)\right) \tag{6}$$

$$KL(f,g) = \sum_{f^*} \sum_{f \in f^*} p(f) KL\left(p(S|f), q(S|f^*)\right). \tag{7}$$

From (6) and (7), we can see that the loss of mutual information can be minimized by minimizing the K–L divergence between $p(F|s)$ and $q(F|s^*)$, as well as the divergence between $p(S|f)$ and $q(S|f^*)$. Thus an iterative co-clustering algorithm can be carried out through the following four steps:

1) ***Initialization***. Assign all auditory scenes and features randomly into k and l clusters, respectively. Then calculate the initial value of the q matrix.
2) ***Update row clusters***. For each row s, find its new cluster index i in the measure of K-L divergence as

$$i = \arg \min_k KL\left(p(F|s), q\left(F|s_k^*\right)\right). \tag{8}$$

Thus the K–L divergence of $p(F|s)$ and $q(F|s^*)$ decreases in this step. With the new row cluster indices, update the q matrix according to (4).
3) ***Update column clusters***. Based on the updated q matrix in step 2, find a new cluster index j for each column f in the measure of K-L divergence, as

$$j = \arg \min_l KL\left(p(S|f), q\left(S|f_l^*\right)\right). \tag{9}$$

Thus the K–L divergence of $p(S|f)$ and $q(S|e^*)$ decreases in this step. With the new column cluster indices, update the q matrix again.
4) ***Re-calculate the loss of mutual information***. If the change in the loss of mutual information is smaller than a predefined threshold [20], stop the iteration process and return the clustering results; otherwise go to step 2 to start a new iteration.

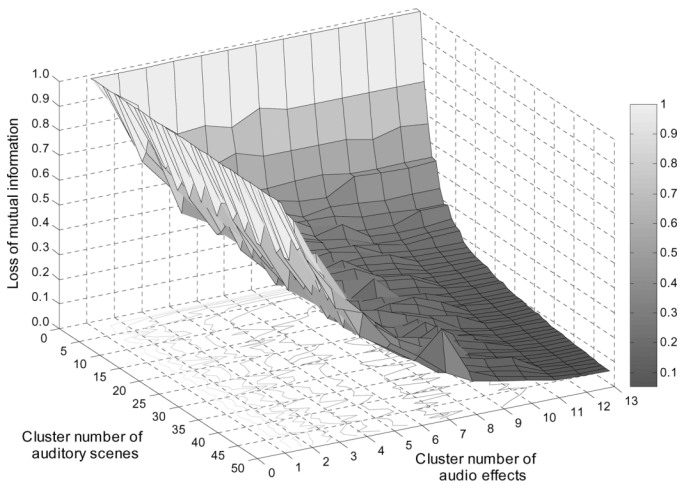

Fig. 3. Illustration of the loss of mutual information with different cluster numbers of auditory scenes and audio effects.

In [19], it has been proved that the above iteration process results in a monotonic decrease in the loss of mutual information and always converges to a local minimum. In the implementation of the process, 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 [20]. The algorithm is computationally efficient and the complexity is $O(n \times \tau \times (k+l))$, where n is the number of nonzeros in $p(S, F)$, and τ is the iteration number.

B. Estimation of the Cluster Numbers

The row and column cluster number k and l, respectively, are assumed to be known in the abovementioned co-clustering algorithm. However, in most applications, it's hard to specify the cluster numbers beforehand, thus an effective approach to automatically estimate the numbers of clusters in co-clustering is highly required.

As mentioned in Section III-A, the criterion used to evaluate the clustering results is the loss of mutual information. However, as Fig. 3 illustrates, the more clusters are used, the more mutual information is kept after the clustering. For example, when both the row and column cluster numbers are one, there is 100% mutual information loss; while there is no mutual information loss if the cluster number is equal to the original sample amount. Thus, we can not get reasonable cluster number if we choose cluster number based on mutual information loss only. However, from the viewpoints of statistics, clusters can be considered as a model describing the data distribution. Therefore, with more clusters, the model complexity (the number of parameters in the model) increases significantly. From this view, we can use the criteria, like Bayesian information criterion (BIC) [26], in selecting the optimal cluster numbers of co-clustering, to balance the loss of mutual information and the model complexity.

BIC has been used successfully to automatically select the cluster number in K-means clustering [27]. 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 co-clustering scheme, the implementation of the BIC criterion is somewhat different from the one frequently used in one-way clustering. First, given the values of k and l, the data likelihood L could be approximated by the logarithm of the ratio between the mutual information after clustering $(I(S^*; F^*))$ and the original mutual information $(I(S; F))$. It is assumed that the model reserving more mutual information would have higher "probability" to fit the data. Second, as co-clustering is a two-way clustering, the model complexity here should consist of two parts: the size of the row clusters (k cluster centers of dimensionality n) and the size of the column clusters (l cluster centers of dimensionality m). Therefore, we give the following definition of the BIC to be used in our co-clustering scheme:

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

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

C. Construction of the Contingency Tables

To employ the co-clustering scheme for auditory scene categorization, the first step is to construct the contingency (or co-occurrence) table. In this section, we show how to construct the contingency tables based on low-level acoustic features and mid-level audio effects, respectively.

1) Acoustic Feature-Based Contingency Table: In ITCC, each cell of the contingency table reflects the co-occurrence probability of corresponding features and scene samples. However, as low-level acoustic features are usually with varying value scopes, they should be normalized prior to the construction of the contingency table.

First, for the jth acoustic feature, its value in the ith auditory scene is re-scaled as according to the following expression:

$$\begin{aligned} \hat{f}_j^i &= \left(f_j^i - f_j^{min}\right) / \left(f_j^{max} - f_j^{min}\right) \\ where \quad f_j^{min} &= \min\left(f_j^i, 1 \le i \le m\right) \\ and \quad f_j^{max} &= \max\left(f_j^i, 1 \le i \le m\right). \end{aligned} \tag{12}$$

Thus, the values of each feature in all the scene samples are in the range of [0, 1]. In our approach, we can adopt such normalized value to approximate the occurrence probability of the related feature in a given auditory scene. For example, the larger the value of the "short-time energy," the higher is the probability of "high volume" in the auditory scenes.

TABLE I
INFORMATION ON THE EXPERIMENTAL AUDIO DATA

Movie/TV Title	Category	Duration
Saving Private Ryan	war	2:41:41
Enemy at the Gates	war	2:11:08
Swordfish	action	1:39:17
The Rock	action	2:16:35
3rd Rock from the Sun	situation comedy	0:30:05
Hollywood Squares	TV shows	0:25:43
59th Annual Golden Globe Awards	TV shows	2:14:50

In the final contingency table $P(S,F)$, each element in $P(S,F)$ is normalized by the sum of all the elements in the table to ensure its integral is one, as

$$p(s_i, f_j) = \frac{\hat{f}_j^i}{\sum_{1\le i\le m}\sum_{1\le j\le n}\hat{f}_j^i}. \tag{13}$$

2) Audio Effect-Based Contingency Table: For each auditory scene, the framework proposed in our previous work [2] is first utilized to detect audio effects. In this framework, hidden Markov models (HMMs) are used to model the audio effects. Then a sliding window of one second moves through the input audio segment with 0.5 s overlap. Each window is further compared against each HMM, which leads to a log-likelihood score (confidence). At last, each window is classified as the audio effect with the highest score.

Based on the detection results, we estimate the occurrence probability of the jth audio effect e_j in the ith auditory scene:

$$p_j^i = \frac{1}{T_i}\sum_{1\le t\le T_i, t\in e_j} c_j(t) \tag{14}$$

where T_i is the number of sliding windows in the ith auditory scene, and $c_j(t)$ is the confidence score of the tth sliding window belonging to audio effect e_j. Then, similar to (13), each element $p(s,e)$ of the contingency table $p(S,E)$ can be calculated as:

$$p(s_i, e_j) = \frac{p_j^i}{\sum_{1\le i\le m}\sum_{1\le j\le n} p_j^i}. \tag{15}$$

IV. EVALUATION

The evaluation of the proposed co-clustering scheme was performed on auditory scene segments extracted from about 12 hours of composite audio selected from various types of video sound tracks, including movies and entertainment TV shows. These sound tracks have abundant audio effects, which are usually distinct enough to be perceived. Detailed information on the audio sequences used in our experiments is listed in Table I. All the audio streams are in 16 KHz, 16-bit and mono channel. Auditory scenes are first labeled by three graduate students and assigned to five semantic categories: *excitement*, *humor*, *pursuit*, *fight*, and *air-attack*. To facilitate further evaluation steps, only scenes with clear semantics are kept. These are the scenes, for which all three labelers were unanimous. In total, 272 scenes were selected into our test set.

TABLE II
CLUSTERING PERFORMANCE ON LOW-LEVEL ACOUSTIC FEATURES WITH SPECIFIED NUMBERS OF CLUSTERS

	No.	*excite.*	*humor*	*pursuit*	*fight*	*air-attack*	preci.
K-means	1	**9**	4	3	1	0	53.33%
	2	**7**	5	1	0	0	
	3	8	**27**	5	2	2	61.36%
	4	2	2	**44**	39	17	42.31%
	5	5	0	36	**47**	6	50.00%
	rec.	51.61%	71.05%	49.44%	52.81%	0.00%	**49.26%**
Co-clustering	1	**23**	6	0	0	0	79.31%
	2	8	**30**	3	3	2	65.22%
	3	0	1	**64**	29	16	58.18%
	4	0	1	21	**41**	4	48.72%
	5	0	0	1	**16**	3	
	rec.	74.19%	78.95%	71.91%	64.04%	0.00%	**63.97%**

The unsupervised clustering results are evaluated against these manually labeled ground truths. To illustrate the efficiency of the co-clustering scheme in auditory scene grouping, we also compare it with some traditional one-way clustering algorithms, as K-means and X-means [27], which are conceptually closest to our approach. Here, the X-means algorithm is an extended K-means algorithm, in which BIC is also used to estimate the cluster number automatically.

The experimental results on low-level acoustic features and on mid-level audio effects are reported in Sections IV-A and IV-B, respectively, and some summarization are also presented in Section IV-C.

A. Co-Clustering of Acoustic Features and Auditory Scenes

Totally, the 29-dimension frame-based acoustic feature vectors are extracted for each auditory scene, including STE, ZCR, brightness, bandwidth, harmonicity prominence [3], subband energy and subband flux [3] (8 subbands), and 8-order MFCCs. In the experiments, the mean values of frame features over an entire auditory scene are taken as the features of that scene.

With the contingency table constructed in Section III-C-1, the evaluations are performed for two cases: 1) assuming the number of clusters is known in *a priori* (compared with K-means); and 2) estimating the numbers of clusters automatically (compared with X-means), respectively.

1) Given the Numbers of Clusters: First, we specify the number of clusters as five in both the K-means and the ITCC algorithm. As the traditional K-means algorithm is sensitive to the choice of initial centroids, we apply the far-apart strategy to select more proper initial centroids for K-means and so to provide more accurate classification results. Although the number of scene clusters is specified, in the co-clustering the

TABLE III
CLUSTERING PERFORMANCE ON LOW-LEVEL ACOUSTIC FEATURES WITH AUTOMATICALLY ESTIMATED NUMBERS OF CLUSTERS

	No.	*excite.*	*humor*	*pursuit*	*fight*	*air-attack*	**preci.**
X-means	1	**8**	3	0	0	0	83.33%
	2	**4**	0	0	0	0	
	3	**6**	1	0	0	0	
	4	**2**	0	0	0	0	
	5	8	**22**	2	1	0	64.86%
	6	2	**2**	0	0	0	
	7	1	2	**45**	34	16	43.65%
	8	0	8	**10**	7	3	
	9	0	0	6	**7**	1	55.29%
	10	0	0	24	**36**	3	
	11	0	0	2	**4**	2	
	rec.	64.52%	63.16%	61.80%	52.81%	0.00%	**53.68%**
Co-clustering	1	**14**	0	0	0	0	92.59%
	2	**11**	2	0	0	0	
	3	5	**32**	2	1	0	80.00%
	4	1	4	**14**	4	6	55.65%
	5	0	0	**32**	14	9	
	6	0	0	**23**	11	6	
	7	0	0	13	**36**	4	72.84%
	8	0	0	5	**23**	0	
	rec.	80.65%	84.21%	77.53%	65.17%	0.00%	**68.01%**

number of feature (column) clusters still needs to be estimated with the BIC based approach, as introduced in Section III-B.

The clustering performance for both algorithms is illustrated in Table II. There, each row represents one obtained cluster, while each column shows the distribution of the scene samples belonging to a ground truth category across all obtained clusters. To make the results even more informative, we shade the table field linking a particular ground truth semantic category and the clusters most likely corresponding to this category, and then calculate the corresponding precision and recall, as well as the average accuracy of the clustering. Here, for each row, the shadow field was automated selected as the one with the largest number. Thus, there may be multiple clusters being associated with one semantic category like the *excitement* and *fight* in Table II; and there also may be some semantic category such as the *air-attack* having no related cluster. This is why the recall of *air-attack* is 0.00% in Table III.[3]

From Table II, it is clear that the co-clustering algorithm yields relatively better performance. For example, the recall of all semantic categories (except for *air-attack*, which is missed by both algorithms) are above 60% using co-clustering, while most of the recalls obtained using K-means are only around 50%. Similar phenomenon can also be observed for the precision. The average accuracy of the co-clustering algorithm is around 64%, while K-means can only reach about 50%. This indicates that the clusters obtained using co-clustering are more distinguishing with respect to the ground truth.

[3]In clustering, samples of *air-attack* were prone to be grouped with samples of *pursuit* because they have confusable acoustic characteristics, like the fourth cluster generated by K-means and the third cluster by co-clustering in Table II. As the number of examples of *air-attack* is small while that of *pursuit* is larger, such mixed clusters are dominated by *pursuit*. Therefore, there is no cluster was associated with the *air-attack*.

TABLE IV
ACOUSTIC FEATURE GROUPS OBTAINED USING ITCC

No.	Features
1	STE, Harmonicity Prominence, $1^{st}-$MFCC, Brightness, $1^{st}-$subband energy, $1^{st}-$subband flux
2	$2^{nd}-$MFCC, $4^{th}-$MFCC $\sim$ $8^{th}-$MFCC
3	ZCR, Bandwidth, $5^{th}-$subband energy $\sim$ $8^{th}-$subband energy, $5^{th}-$subband flux $\sim$ $8^{th}-$subband flux
4	$3^{rd}-$MFCC, $2^{nd}-$subband energy $\sim$ $4^{th}-$subband energy, $2^{nd}-$subband flux $\sim$ $4^{th}-$subband flux

2) Automatic Estimation the Number of Clusters: We also evaluate the BIC-based strategy proposed in Section III-B for automatically selecting the cluster numbers in the co-clustering scheme, and—for comparison—also in the X-means algorithm. In the X-means clustering, we search for the proper cluster number K in the range of $(1 \leq K \leq 50)$, and in the co-clustering, we look for the auditory scene cluster number k and the acoustic feature cluster number l in the ranges of $1 \leq k \leq 50$ and $1 \leq l \leq 29$.

Table III shows the clustering performance using the estimated numbers of clusters. Finally, we get 11 and eight auditory scene categories with the X-means and ITCC algorithms, respectively. Table III shows that compared to the ITCC algorithm, the X-means algorithm results in a stronger over-segmentation of data with respect to the ground truth categories. For example, four clusters are found that are related to the *excitement* category, while only two such clusters are obtained using ITCC. In other words, the cluster structure obtained using ITCC resembles more closely the content distribution across the defined semantic categories. Moreover, the average classification accuracy of ITCC is close to 70%, while that of X-means is only just above 50%.

In Table IV, we list the clusters of acoustic features resulting from co-clustering. There are four feature groups, and the grouping results look reasonable regarding the nature of the features. For example, most of the MFCC coefficients are grouped together in the second cluster, subband features in high frequency domain are classified in the third cluster, while subband features in low frequency range are mainly in the first and the fourth clusters.

B. Co-Clustering of Audio Effect and Auditory Scenes

For auditory scene categorization based on mid-level audio effects, the HMM-based framework proposed in our previous work [2], [3] was utilized to detect ten audio effects including *applause*, *car racing*, *cheer*, *car-crash*, *explosion*, *gun-shot*, *helicopter*, *laughter*, *plane*, and *siren*, as well as three general audio effects, such as *music*, *speech*, and *noise*, in all the scene samples. These audio effects were selected in order to represent the five auditory scene categories defined in our experiments.

TABLE V
CLUSTERING PERFORMANCE ON MID-LEVEL AUDIO EFFECTS WITH SPECIFIED NUMBERS OF CLUSTERS

	No.	*excite.*	*humor*	*pursuit*	*fight*	*air-attack*	**preci.**
K-means	1	31	**38**	3	1	0	52.05%
	2	0	0	**43**	9	23	69.09%
	3	0	0	**33**	2	0	
	4	0	0	7	**60**	0	86.52%
	5	0	0	3	**17**	2	
	rec.	0.00%	100%	85.39%	86.52%	0.00%	**70.59%**
Co-clustering	1	**17**	0	1	0	0	96.30%
	2	**9**	0	0	0	0	
	3	5	**38**	0	0	0	88.37%
	4	0	0	**80**	6	25	72.07%
	5	0	0	8	**83**	0	91.21%
	rec.	83.87%	100%	89.89%	93.26%	0.00%	**83.46%**

The general precision and recall of audio effect detection can reach around 80% and 90%, respectively. Moreover, the detection errors mostly occur among audio effects usually happening together. For example, *gun-shot* and *explosion* are found to be easily misclassified as one being the other and vice verse. The evaluation results presented below show that our co-clustering scheme suffers a little from such misclassification among co-occurring audio effects. Similar to the evaluation steps in Section IV-A, the categorization process is performed here also for two cases, i.e., for a given number of clusters, and with automatically estimated clusters.

1) Given Number of Clusters: Here, we also prespecify the number of cluster as five for both K-means and ITCC. The clustering results are shown in Table V.

Table V shows that the ITCC-based approach again achieved better performance than K-means. When using K-means, two semantic categories, i.e., *excitement* and *air-attack*, were missed. For example, most samples from *excitement* are mixed with samples from *humor* in the obtained cluster one. Compared to this, samples from *excitement* and *humor* are correctly distinguished by co-clustering. The general classification accuracy of ITCC is also evidently higher than K-means.

2) Automatic Estimation the Number of Clusters: In the X-means clustering, we search for the proper cluster number K in the range of $1 \leq K \leq 50$; and in ITCC, we look for the scene cluster number k and the effect cluster number l in the ranges of $1 \leq k \leq 50$, and $1 \leq l \leq 13$, respectively.

In our experiments, we finally obtained eight auditory scene categories by using the ITCC, and 13 scene categories by using the X-means clustering, as that shown in Table VI. This table shows that, in general, the co-clustering algorithm can achieve better performance in the auditory scene categorization. First, the number of auditory categories obtained by co-clustering is closer to the ground truth than that achieved with the X-means clustering. This indicates that co-clustering can give a more exact approximation of the actual content grouping structure in the test database. Second, for most categories, co-clustering achieves a higher precision and recall than the X-means algorithm. In average, around 88.6% auditory scenes are correctly classified with the co-clustering algorithm, and the performance of the X-means clustering is 83.82%.

TABLE VI
CLUSTERING PERFORMANCE ON MID-LEVEL AUDIO EFFECTS WITH AUTOMATICALLY ESTIMATED NUMBERS OF CLUSTERS

	No.	*excite.*	*humor*	*pursuit*	*fight*	*air-attack*	**preci.**
X-means	1	**21**	0	0	0	0	100%
	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**	
	rec.	67.74%	100%	74.16%	91.01%	88.00%	**83.82%**
Co-clustering	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%
	rec.	87.10%	100%	78.65%	91.01%	100%	**88.60%**

TABLE VII
AUDIO EFFECT GROUPS OBTAINED USING ITCC

No.	**Audio Effects**	**No.**	**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*	–	

Also, with the co-clustering algorithm we obtain seven rather intuitive audio effect groups, as illustrated in Table VII. One interesting case to note in Table VII is that *music* is grouped together with *helicopter* and *siren*. After investigating the audio tracks we found that in action movies most *pursuit* scenes, in which sounds of *helicopter* and *siren* frequently occur, are usually accompanied by a *music* background.

C. Summary of the Evaluation Results

In the above experiments, we have evaluated the proposed co-clustering scheme for auditory scene categorization through its comparison with the traditional one-way K-means (X-means) algorithms, using both low-level acoustic features and mid-level audio effects, and using both pre-specified

number of clusters and automatically estimated cluster numbers. In view of the detailed explanations of the results we gave in Sections IV-A and IV-B, we can summarize as follows.

- In general, the two-way co-clustering scheme outperforms traditional one-way clustering, both when using the low-level acoustic features and mid-level audio effects, and also both with given number of clusters and auto-estimated clusters. The clusters obtained with co-clustering are more consistent with the ground truth semantic categories, thus its average clustering accuracy is higher than K-means/X-means.
- The proposed BIC-based strategy is effective in estimating the numbers of clusters in co-clustering, on both low-level and mid-level feature spaces. With this strategy, we achieved the number of clusters which is closer to the ground truth, compared to the X-means algorithm (see Table III and Table VI).
- Compared with low-level features, mid-level audio effects can provide more comprehensive descriptions to auditory scenes and achieve better grouping results, using both K-means and ITCC. Such phenomenon is consistent with the results obtained in previous research works on auditory scene classification with supervised statistical learning [3]–[5].
- The co-clustering scheme is more robust than other schemes for both low-level and mid-level based scene categorization. In comparison, with K-means, the recall on some clusters in the mid-level feature space is even worse than in the low-level feature space. For example, two semantic categories (*excitement* and *air-attack*) were missed by K-means clustering in the mid-level space (Table V) while only one (*air-attack*) was missed in the low-level space (Table II). In comparison, the co-clustering is more stable.
- The best performance in the whole experiment was obtained using the co-clustering scheme and the mid-level audio effects.

V. Discussions

The results show that the proposed co-clustering scheme achieves better performance on clustering both the low-level acoustic features and mid-level audio effects. In view of the fact that we worked with general audio features and treated all audio data in the same way, independent of various possible combinations of *speech*, *music*, and *noise*, these results indicate that our approach can be employed widely, in various audio content analysis applications. In pervasive computing scenarios, where any low-level acoustic features can be extracted, our scheme can be performed directly on low-level feature spaces to assign audio segments to semantic categories. In some other applications, where typical audio effects can be enumerated and detected, our approach can achieve better performance on such mid-level representations. Moreover, to further improve the semantic categorization results, it is still possible to expand our scheme with other information such as speech transcription [28], and with other technologies such as temporal sequence modeling.

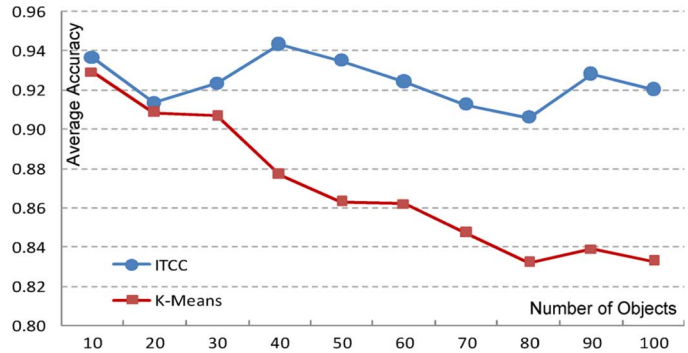

Fig. 4. Comparison of ITCC and K-means on image clustering. The clustering performances with regard to various number of image objects are examined. The ground truth is from the COIL-100 database.

More generally, we believe our approach can also facilitate other multimedia applications. As we mentioned in the section of Introduction, to provide more general solutions and move away from those narrow scope approaches [13], more and more recent research works have pay attention to develop unsupervised solutions for multimedia content analysis. As a basic technology, unsupervised clustering acts an important role in building these general solutions. In this paper, we have investigated the fundamental problem—finding proper feature correlations and similarity measurement—in clustering, and proposed ITCC as a feasible scheme to formulize this problem. To the best of our knowledge, only the manifold-based algorithms (such as the spectral clustering [29]) impliedly consider such problem in clustering. However, in comparison to manifold-based algorithms, our approach provide another complete different solution, and can explicitly discover feature groups (as those shown in Tables IV and VII), which may also benefit some applications. We have successfully applied our approach in audio analysis in this paper, and our experience may benefit attempts in other multimedia modals.

To show the effectiveness of the co-clustering scheme on other mediums, in this paper we also carried out a preliminary evaluation on content-based image clustering, which is one of the fundamental components in image related applications. The evaluation was based on the well-known COIL-100 database which consists of 7200 images from 100 objects, with 72 images for each object [30]. Further, to examine the clustering performance of the proposed scheme with regard to various number of image objects, we randomly selected ten, 20, ..., and 100 objects from the COIL-100 database, and then applied clustering on these selected objects. The feature used in the experiment was the 64-dimentional histogram in HSV color space [31], which is widely used in image retrieval. Similar to the experiments in Section IV, we also compared our ITCC method with K-means, and use accuracy as the criterion for evaluation. The performance is shown in Fig. 4. From Fig. 4, it's clear that ITCC always achieved better accuracy than K-means. This result is consistent with the previous experiments on auditory categorization in Section IV. Moreover, with the increasing of the number of objects in the clustering, the performance of K-means drops significantly, while that of ITCC is relatively stable. More objects here means more images are used, and

more possibilities of misclassification. This is why the accuracy of K-means drops when clustering more objects. In comparison, the stable accuracy of ITCC indicates it is more robust in dealing with a large number of objects.

VI. Conclusion

In this paper, we focused on the problem of automatically categorizing auditory scene in an unsupervised manner. We first investigated the phenomenon of local grouping trends among various low- and mid-level features, and explained how such trends will affect the semantic similarity measurements between scenes. An approach via co-clustering proved to be capable to exploit grouping trends among different dimensions of the feature space, and provide semantically the most accurate way to measure similarity between auditory scenes. Moreover, we extended the algorithm with a BIC-based strategy to automatically select the numbers of clusters in co-clustering. Experiments showed encouraging results on both semantic categorization and cluster number estimation. Co-clustering achieved better performances in comparison with traditional one-way clustering, on both the low-level acoustic features and the mid-level audio effect representations.

There is still room to improve the proposed scheme. For example, in the current scheme the contingency table must be re-computed when new samples are added. This practical inconvenience may be addressed by generalizing the proposed co-clustering process towards an incremental clustering scheme. Moreover, in co-clustering, it is still challenging to model temporal relationship among various audio effects to improve the clustering performance. This is also a good direction in the further work.

References

[1] H. Sundaram and S.-F. Chang, "Audio scene segmentation using multiple features, models and time scales," in *Proc. IEEE ICASSP'00*, Istanbul, Turkey, Jun. 2000, vol. 4, pp. 2441–2444.

[2] R. Cai, L. Lu, H.-J. Zhang, and L.-H. Cai, "Highlight sound effects detection in audio stream," in *Proc. IEEE ICME'03*, Baltimore, MD, Jul. 2003, vol. 3, pp. 37–40.

[3] R. Cai, L. Lu, A. Hanjalic, H.-J. Zhang, and L.-H. Cai, "A flexible framework for key audio effects detection and auditory context inference," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 3, pp. 1026–1038, May 2006.

[4] M. Xu, N. Maddage, C.-S. Xu, M. Kankanhalli, and Q. Tian, "Creating audio keywords for event detection in soccer video," in *Proc. IEEE ICME'03*, Baltimore, MD, Jul. 2003, vol. 2, pp. 281–284.

[5] W.-H. Cheng, W.-T. Chu, and J.-L. Wu, "Semantic context detection based on hierarchical audio models," in *Proc. ACM MIR'03*, CA, USA, Nov. 2003, pp. 109–115.

[6] W.-T. Chu, W.-H. Cheng, and J.-L. Wu, "Generative and discriminative modeling toward semantic context detection in audio tracks," in *Proc. IEEE MMM'05*, Melbourne, Australia, Jul. 2005, pp. 38–45.

[7] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," in *Proc. IEEE ICASSP'02*, Orlando, USA, May 2002, vol. 2, pp. 1941–1944.

[8] A. Eronen, J. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context awareness—Acoustic modeling and perceptual evaluation," in *Proc. IEEE ICASSP'03*, Hong Kong, Apr. 2003, vol. 5, pp. 529–532.

[9] A. Eronen, V. Peltonen, J. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Trans. Audio Speech Lang.*, vol. 14, no. 1, pp. 321–328, Jan. 2006.

[10] L. Ma, D. Smith, and B. Milner, "Context awareness using environmental noise classification," in *Proc. EUROSPEECH'03*, Geneva, Switzerland, Sep. 2003, pp. 2237–2240.

[11] D. P. W. Ellis and K. Lee, "Minimal-impact audio-based personal archives," in *Proc. ACM Workshop on Continuous Archival and Retrieval of Personal Experience*, New York, Nov. 2004, pp. 39–47.

[12] N. Kern, A. Schmidt, and B. Schiele, "Recognizing context for annotating a live life recording," *Personal Ubiquitous Comput.*, vol. 11, no. 4, pp. 251–263, 2007.

[13] A. Hanjalic, J. Nesvadba, and J. Benois-Pineau, "Moving away from narrow scope solutions in multimedia content analysis," in *Proc. IEE EWIMT'05*, London, U.K., 2005.

[14] B. Clarkson and A. Pentland, "Unsupervised clustering of ambulatory audio and video," in *Proc. IEEE ICASSP'99*, AZ, Mar. 1999, vol. 6, pp. 3037–3040.

[15] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[16] D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer, "Co-clustering of biological networks and gene expression data," *Bioinformatics*, vol. 18, no. 90001, pp. S145–S154, 2002.

[17] S. Madeira and A. Oliveira, "Biclustering algorithms for biological data analysis: A survey," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 1, no. 1, pp. 24–45, 2004.

[18] I. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *Proc. ACM SIGKDD'01*, San Francisco, CA, Aug. 2001, pp. 269–274.

[19] I. Dhillon, S. Mallela, and D. Modha, "Information-theoretic co-clustering," in *Proc. ACM SIGKDD'03*, Washington, DC, USA, Aug. 2003, pp. 89–98.

[20] I. Dhillon and Y. Guan, "Information theoretic clustering of sparse co-occurrence data," in *Proc. IEEE ICDM'03*, Melbourne, FL, Nov. 2003, pp. 517–520.

[21] N. Friedman, "The Bayesian structure EM algorithm," in *Proc. UAI'98*, Madison, WI, Jul. 1998, pp. 129–138.

[22] J. Shlens, A Tutorial on Principal Component Analysis Dec. 2005 [Online]. Available: http://www.snl.salk.edu/~shlens/pub/notes/pca.pdf

[23] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, no. 4–5, pp. 411–430, 2000.

[24] R. El-Yaniv and O. Souroujon, "Iterative double clustering for unsupervised and semi-supervised learning," in *Proc. ECML'01*, Freiburg, DE, Sep. 2001, pp. 121–132.

[25] J. Hartigan, "Direct clustering of a data matrix," *J. Amer. Stat. Assoc.*, vol. 67, no. 337, pp. 123–129, 1972.

[26] R. Kass and L. Wasserman, "A reference Bayesian test for nested hypotheses and its relationship to the schwarz criterion," *J. Amer. Staistt. Assoc.*, vol. 90, no. 431, pp. 928–934, 1995.

[27] D. Pelleg and A. Moore, "X-means: Extending K-means with efficient estimation of the number of clusters," in *Proc. ICML'00*, Stanford, CA, USA, Jun. 2000, pp. 727–734.

[28] D. K. Roy, "Learning visually-grounded words and syntax for a scene description task," *Comput. Speech Lang.*, vol. 16, no. 3–4, pp. 353–385, 2002.

[29] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. NIPS'01*, 2001, pp. 849–856.

[30] S. A. Nene, S. K. Nayar, and H. Murase, Columbia Object Image Library (COIL-100) Tech. Rep. CUCS-006-96, Feb. 1996 [Online]. Available: http://www1.cs.columbia.edu/CAVE/software/softlib/coil-100.php

[31] S. Sural, G. Qian, and S. Pramanik, "Segmentation and histogram generation using the HSV color space for image retrieval," in *Proc. IEEE ICIP'02*, Rochester, New York, 2002, vol. 2, pp. 589–592.

Rui Cai received the B.E. and Ph.D. degrees from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2001 and 2006, respectively.

Since 2006, he has been with Microsoft Research Asia, Beijing, where he is currently an Associate Researcher with the Web Search and Mining Group. His current research interests include web search and data mining, machine learning and pattern recognition, multimedia content analysis, and signal processing.

Dr. Cai is a member of ACM. He served as technical program committee member for International Conference on Distributed Computing Systems 2008 and Asia Information Retrieval Symposium 2008.

Lie Lu (M'05) received the B.S. and M.S. degrees from Shanghai Jiao Tong University, China, both in electrical engineering, in 2000 and 1997, respectively. He is currently pursuing Ph.D. degree from Delft University of Technology, The Netherlands.

Since 2000, he has been with Microsoft Research Asia, Beijing, where he is currently a Researcher with the Speech Group. His current research interests include pattern recognition, content-based audio analysis and indexing, and content-based music analysis. He has authored about 50 publications in these areas and has 15 patents or pending applications.

Mr. Lu is a member of the ACM. He also served as a member of technical program committee for several conferences, such as the IEEE International Conference on Multimedia and Expo 2004 and 2007, IJCAI Workshop on Multimodal Information Retrieval 2007, ACM Multimedia 2007, Pacific-Rim Conference on Multimedia 2007, and Asia Information Retrieval Symposium 2008.

Alan Hanjalic (M'00) received the Dipl.-Ing. degree in 1995 from the Friedrich-Alexander University in Erlangen, Germany, and the Ph.D. degree in 1999 from the Delft University of Technology, both in electrical engineering.

He is an Associate Professor of Multimedia Technology, and Head of the Multimedia Content Analysis research cluster at the Department of Mediamatics of the Delft University of Technology, The Netherlands. He was a Visiting Scientist at Hewlett-Packard Labs, Palo Alto CA, a Research Fellow at British Telecom Labs, Ipswich, U.K., and a Visiting Scientist at Philips Research Labs, Briarcliff Manor NY, as well as at Microsoft Research Asia, Beijing, China. His research interests and expertise are in the broad areas of multimedia signal processing, media informatics, and multimedia information retrieval, with focus on multimedia content analysis for interactive content browsing and retrieval and on personalized and on-demand multimedia content delivery. In his areas of expertise he authored and co-authored more than 70 publications, among which the books titled *Image and Video Databases: Restoration, Watermarking and Retrieval* (Elsevier, 2000) and *Content-Based Analysis of Digital Video* (Kluwer, 2004), and holds several EU/US patents.

Dr. Hanjalic is a member of Editorial Boards of the IEEE TRANSACTIONS ON MULTIMEDIA and *Journal of Multimedia*. He is a Guest Editor of the *International Journal of Image and Graphics* (World Scientific, Singapore), Special Issue on Content-based Image and Video Retrieval, July 2001, and of the IEEE TRANSACTIONS ON MULTIMEDIA, Special Issue on Integration of Context and Content for Multimedia Management, 2008. He also serves as the Chief Guest Editor of the PROCEEDINGS OF THE IEEE, Special Issue on Advances in Multimedia Information Retrieval, 2008. He was a Chair of the IS&T/SPIE Conferences on Multimedia Content Analysis, Management and Retrieval 2006, and on Multimedia Content Access: Algorithms and Systems 2007, Program Chair (Content Track) of the ACM Multimedia Conference 2007, Workshops Chair of the ACM Multimedia Conference 2006, and as Area/Track Chair of numerous conferences, including the Multimedia Modeling Conference 2007, the IEEE ICME 2007 conference, and the Pacific Rim Conference on Multimedia 2007.