2004 IEEE International Conference on Multimedia and Expo (ICME)

A Study of Semantic Context Detection by Using SVM and GMM Approaches

Wei-Ta Chu, Wen-Huang Cheng, Ja-Ling Wu, and Jane Yung-jen Hsu[†]
Department of Computer Science and Information Engineering
National Taiwan University
No. 1, Sec. 4, Roosevelt Road, Taipei, Taiwan 106
{wtchu, wisley, wjl}@cmlab.csie.ntu.edu.tw, [†] yjhsu@csie.ntu.edu.tw

Abstract

Semantic-level content analysis is a crucial issue to achieve efficient content retrieval and management. In this paper, we propose a hierarchical approach that models the statistical characteristics of several audio events over a time series to accomplish semantic context detection. Two stages, including audio event and semantic context modeling/testing, are devised to bridge the semantic gap between physical audio features and semantic concepts. HMMs are used to model audio events, and SVMs and GMMs are used to fuse the characteristics of various audio events related to some specific semantic concepts. The experimental results show that the approach is effective in detecting semantic context. The comparison between SVM- and GMM-based approaches is also studied.

1. Introduction

With the rapid development of technologies in content creation, storage, and dissemination, tremendous amount of multimedia digital contents have been created and applied in many fields. However, the massive multimedia data, including video, audio, and text, impede users in content browsing, retrieval, and management. Many techniques of content analysis were proposed to facilitate content management in recent years. For example, digital content is classified and segmented by analyzing audio features in [1] and [2]. Furthermore, techniques on analyzing various types of video data were also proposed to facilitate content management [3, 4].

Although the studies described above effectively achieve data classification and segmentation, the techniques which only consider data features still don't meet users' needs. An approach which takes users' sense into account is necessary to provide semantic-level content retrieval/management. In [5], an approach based on HMM was proposed to detect highlight sound effects such as applause, cheer, and laughter in audio streams. The results of audio event detection provide the clues for exploring the semantics of a video segment. However, users would more likely to find a scene which possesses a complete semantic meaning rather than some specific audio/video events. For example, in an action movie, we would like to find the scene of gunplay, which may consist of gunshots, explosions, sounds of jeeps, and screams from soldiers for a while. Therefore, in this paper, we propose a hierarchical approach that models high-level audio scenes based on the results of audio events detection.

We define a 'semantic context' as a complete scene which possesses a single meaning over a time series. For example, the gunplay scenes in an action movie. In this work, we would like to bridge the semantic gap between physical audio features and semantic contexts. Two approaches (SVM and GMM) that model high-level scenes based on audio event detection are proposed and compared.

The rest of this paper is organized as follows: Section 2 describes the overall system framework. The method of audio event modeling and detection is stated in Section 3. Two approaches, including SVM and GMM, are described in Section 4. Section 5 shows the experimental results of semantic context detection, and Section 6 gives the concluding remarks.

2. System framework

The proposed system consists of two stages: audio event detection and semantic context detection. First, as shown in Figure 1(a), the input audio stream is divided into overlapped segments, and several features are extracted from each segment. For each audio event, an HMM is constructed to model extracted features. Through the Viterbi algorithm, the probability for each audio event is computed. The confidence score which describes how likely a segment belongs to an audio event is obtained by using a soft decision strategy. We

0-7803-8603-5/04/$20.00 ©2004 IEEE

say that the segments with high confidence scores from the gunshot model, for example, represent the occurrences of gunshot events.

At the stage of semantic context detection, as shown in Figure 1(b), the characteristics of the confidence scores obtained in the first stage are extracted and modeled by SVMs or GMMs. Detailed descriptions about modeling and detection will be given in the following sections.

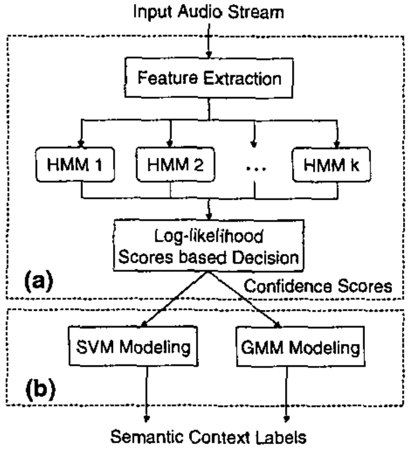

Figure 1. The proposed system framework contains (a) audio event and (b) semantic context detection

3. Audio event modeling

Audio events in this paper are defined as short audio segments which represent the sound of an object or an event, e.g. gunshots, explosion, laughter, etc. The results of audio event detection provide the clues for recognizing higher level semantics of an audio clip.

For modeling two semantic contexts, i.e. *gunplay* and *car-chasing*, in action movies, the audio events we modeled are *gunshots*, *explosions*, *helicopter-flying*, *engines*, and *car-braking*. In this work, all audio clips are down-sampled to 16 kHz, 16 bits and mono-channel format. For each audio event, 30 short audio clips each with length 3-10 sec are selected from various action movies as the training data. In the training stage, as shown in Figure 1(a), audio features are extracted from the training data and are modeled by HMMs. In our work [7], input audio streams are first divided into 1-sec segments with 0.5-sec overlapping. Each segment is further segmented into 25-ms audio frames with 10-ms overlapping. The audio features extracted from each frame are volume, band energy ratio, zero-crossing rate, frequency centroid, bandwidth, and 8-order MFCC [6]. For each audio event, an HMM is constructed with 4 states (the model size is estimated by the adaptive sample set construction technique [5]), and the observation probability in each state is described as a 4-mixture Gaussian distribution. After HMM modeling, two distributions, say within and outside distributions, are computed for each audio event. They describe the score distributions of samples within and/or outside an audio event.

In the test stage, audio features from test data are input to all HMMs, and the likelihood with respect to each audio event is obtained. Note that we cannot simply classify an audio segment to a specific event even if its likelihood value is larger than that of other events. A data segment may not belong to any audio event. Therefore, the Neyman-Person test is applied to compute confidence scores, which represent the probability in one class respect to others. Assume that the log-likelihood value is x, then $f_X(x|\theta_1)$ and $f_X(x|\theta_0)$ denote the corresponding probabilities of the within and the outside distributions, respectively. According to the Neyman-Pearson theory, the likelihood ratio can be applied to determine the confidence scores as follows:

$$s(x)=\frac{f_X(x|\theta_1)}{f_X(x|\theta_0)} \tag{1}$$

Through these procedures, the confidence score for each audio segment can be obtained.

4. Semantic context modeling

We aim at detecting high-level semantic context based on the results of audio event detection described in the previous section. To characterize a semantic context, the confidence scores for some specific audio events, which are highly relevant to the semantic concept, are collected and modeled. In our work, the scores from *gunshot*, *explosion*, and *helicopter* events are used to characterize 'gunplay' scenes. The scores from *engine* and *car-braking* events are used to characterize 'car-chasing' scenes.

Note that an audio scene may not contain all relevant audio events at every time instant. For example, in Figure 2, the audio clip from t1 to t2 is a typical gunplay scene which contains mixed relevant audio events. In contrast to this case, no relevant event exists from t4 to t5 and t6 to t7. However, the whole audio segment from t3 to t8 is viewed as a single scene in users' sense, as long as the duration of the 'irrelevant clip' doesn't exceed a threshold. Therefore, to model the characteristics of semantic contexts, we propose two approaches based on SVM and GMM to fuse the information obtained from low-level audio events detection.

In the semantic context modeling, the manually labeled training data which are complete audio scenes (like the segment from t3 to t8 in Figure 2) are input to the audio event detection module. After that, as described in Section 3, the confidence scores of each 1-sec segment are obtained. These segments are called *analysis windows* and are viewed as units of audio events. In order to characterize the long term nature of sound 'texture', we calculate means and variances over a number of *analysis windows*. As shown in Figure 3, the overlapped '*texture windows*' with length 5 sec describe the characteristic of an audio clip.

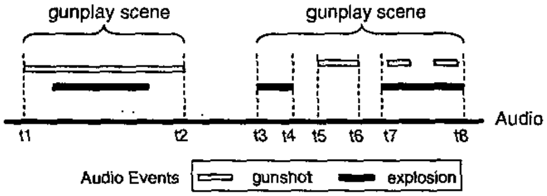

Figure 2. Examples of audio semantic contexts

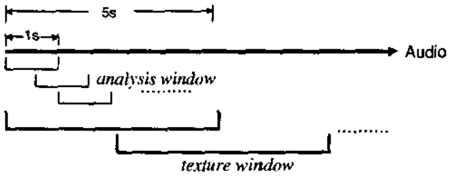

Figure 3. Analysis window and texture window in semantic context modeling

4.1 SVM-based modeling

We exploit SVM classifiers to distinguish the textures of 'gunplay', 'car-chasing', and 'others' scenes. For each texture window, the means and variances of confidence scores obtained from audio event detection are concatenated as a vector. The vectors of an audio clip are then collected as the data for SVM training. According to the performance analysis of multiclass SVM classifiers [8], we apply the 'one-against-one' strategy to model these three scenes. Three SVM models are constructed, i.e. 'gunplay vs. car-chasing', 'gunplay vs. others', and 'car-chasing vs. others'. For each model, the radical basis function (RBF) is used to map features into a higher dimensional space:

$$K(x,y) = \exp(-\gamma \left\| x_i - x_j \right\|^2), \quad \gamma > 0, \qquad (2)$$

where x_i and x_j are the feature vectors of texture windows and γ is the kernel parameter.

In the test stage of semantic context detection, the Decision Directed Acyclic Graph strategy [9] is applied to perform multiclass classification. Figure 4 shows the testing process. The vectors from test data are first input to the root SVM classifier, i.e. 'car-chasing vs. others' classifier, in this case. After this evaluation, the process branches to left if more vectors are predicted as 'others' segments than 'car-chasing' segments. The 'gunplay vs. others' classifier is then used to re-evaluate the testing vectors. After these two steps, the vectors which represent 5-sec audio segments (texture windows) are labeled as 'gunplay' or 'others' scenes.

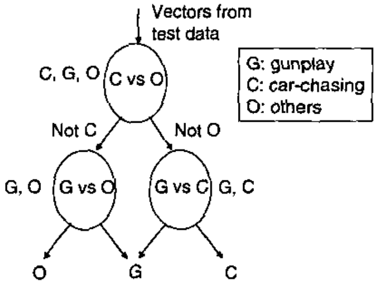

Figure 4. Test process of multiclass SVM

4.2 GMM-based modeling

A GMM-based approach is also applied to fuse the information of audio events detection [7]. As described in Section 3, the confidence scores for each analysis window are calculated. If the scores are larger than a pre-defined threshold, we say that this 1-sec segment belongs to a specific audio event, i.e. one 'occurrence' exists at this time. Unlike the features for SVM training, the characteristic of a texture window is obtained by calculating the occurrence ratio of one audio event. Because the overlapping factor is 0.5, a 5-sec texture window contains nine analysis windows. If four of the nine segments are detected as gunshots audio events, the occurrence ratio of gunshots in this texture window is 0.444. The occurrence ratios collected from all texture windows are then modeled by a multivariate Gaussian distribution.

In the test stage, the occurrence ratios are calculated by applying the same process as the training stage. For each texture window, if all ratio numbers of relevant events (e.g. engine and car-braking events for car-chasing scenes) are larger than the mean values of corresponding Gaussian mixtures, we say that the texture window belongs to the specific semantic context.

5. Experiments

The training and testing audio data of 'gunplay' and 'car-chasing' scenes are manually collected from Hollywood action movies. The audio scenes 'others' contain segments from various video data other than action movies, including music videos, news broadcast, sports games, and soap operas. The total training data for each audio event is about 4 minutes long, and that for each semantic context is about 20 minutes. Note that the criteria of selecting training data for audio events and semantic contexts are different. For semantic context modeling, we collected the gunplay and car-chasing scenes based on the experienced users' subjective judgments, no matter how many relevant audio events exist in the scene. On the contrary, the training data for audio event modeling are many short audio segments that are exactly the audio events. For each semantic context, several audio clips with total length 50 minutes are extracted from several movies as testing data.

We compare the performance of two proposed approaches. Table 1 shows the performance of semantic context detection. In average, the SVM-based approach has better recall performance and similar precision performance with respect to the GMM-based approach. This result meets the requirements for video summarization and retrieval.

Actually, the performance of semantic context detection is data-dependent. The results are affected by different acoustic conditions and the scene textures controlled by the movie directors. Table 2 shows the robustness of detection performance based on SVM and GMM approaches. The results indicate that the SVM-based approach is more robust than GMM. The reason may be that the feature values modeled by GMMs are too sensitive to the variations of different test data. Moreover, the modeling ability of a GMM depends on the number of mixtures it contains, but this number may vary in different testing data.

Table 1. Performance of semantic context detection

Semantic Context		Recall	Precision	False Alarm
Gunplay	SVM	0.798	0.715	0.326
	GMM	0.658	0.670	0.386
Car-chasing	SVM	0.651	0.829	0.197
	GMM	0.570	0.887	0.128

Table 2. Robustness of detection performance

Semantic Context		Var. of Recall	Var. of Precision
Gunplay	SVM	0.002	0.029
	GMM	0.033	0.031
Car-chasing	SVM	0.021	0.022
	GMM	0.045	0.012

Overall, we have better performance on 'gunplay' scene detection. It is believed that 'gunplay' scenes have stronger sound effects and steadier patterns than that in 'car-chasing' scenes. Furthermore, the audio events chose to model car-chasing scenes may be changed or added.

6. Conclusion

In this paper, we have presented a hierarchical framework for semantic context detection. Two audio scenes in action movies, i.e. gunplay and car-chasing, are considered in this work. HMMs are used to model five audio events, and SVM- and GMM-based fusion schemes are developed to characterize high-level audio scenes. Experimental evaluations have shown that the SVM-based approach is more robust than GMM and obtain satisfying results in recall and precision performance. This framework could be applied to various semantic contexts and is flexible to enhance the feasibility by taking video objects/events into account.

7. References

[1] T. Zhang and C.-C.J. Kuo, "Hierarchical System for Content-based Audio Classification and Retrieval", Proceedings of SPIE, Multimedia Storage and Archiving System III, Vol. 3527, pp. 398-409, 1998.

[2] L. Lu, H.-J. Zhang, and H. Jiang, "Content Analysis for Audio Classification and Segmentation", IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 7, pp. 504-516, 2002.

[3] D. Gatica-Perez, A. Loui, and M.-T. Sun, "Finding Structure in Home Videos by Probabilistic Hierarchical Clustering", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 13, No. 6, pp. 539-548, 2003.

[4] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic Soccer Video Analysis and Summarization", IEEE Transactions on Image Processing, Vol. 12, No. 7, pp. 796-807, 2003.

[5] R. Cai, L. Lu, H.-J. Zhang, and L.-H. Cai, "Highlight Sound Effects Detection in Audio Stream", Proceedings of IEEE International Conference on Multimedia & Expo, Vol. 3, pp. 37-40, 2003.

[6] Y. Wang, Z. Liu, and J.-C. Huang, "Multimedia Content Analysis Using Both Audio and Video Clues", IEEE Signal Processing Magazine, Vol. 17, pp. 12-36, 2000.

[7] W.-H. Cheng, W.-T. Chu, and J.-L. Wu, "Semantic Context Detection based on Hierarchical Audio Models", Proceedings of International Workshop on Multimedia Information Retrieval, pp. 109-115, 2003.

[8] C.-W. Hsu and C.-J. Lin, "A Comparison of Methods for Multiclass Support Vector Machines", IEEE Transactions on Neural Networks, Vol. 13, No. 2, pp. 415-425, 2002.

[9] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large Margin DAGs for Multiclass Classification", in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, Vol. 12, pp. 547-553, 2000.