A FUSION SCHEME OF VISUAL AND AUDITORY MODALITIES FOR EVENT DETECTION IN SPORTS VIDEO

Min Xu, Ling-Yu Duan, Chang-Sheng Xu, Qi Tian

Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
{xumin, lingyu, xucs, tian}@lit.org.sg

ABSTRACT

In this paper, we propose an effective fusion scheme of visual and auditory modalities to detect events in sports video. The proposed scheme is built upon semantic shot classification, where we classify video shots into several major or interesting classes, each of which has clear semantic meanings. Among major shot classes we perform classification of the different auditory signal segments (i.e. silence, hitting ball, applause, commentator speech) with the goal of detecting events with strong semantic meaning. For instance, for tennis video, we have identified five interesting events: serve, reserve, ace, return, and score. Since we have developed a unified framework for semantic shot classification in sports videos and a set of audio mid-level representation with supervised learning methods, the proposed fusion scheme can be easily adapted to a new sports game. We are extending this fusion scheme to three additional typical sports videos: basketball, volleyball and soccer. Correctly detected sports video events will greatly facilitate further structural and temporal analysis, such as sports video skimming, table of content, etc.

1. INTRODUCTION

Humans tend to use high-level semantic concepts when querying and browsing multimedia databases; consequently, semantic video indexing is becoming an important task of video documents analysis. For building efficient video management, it is critical to bridge the semantic gap between the simplicity of available visual and auditory features and the richness of user semantics. To address this issue, most existing approaches focus on the integration and evaluation of domain specific knowledge and pattern recognition techniques. For example, D. Zhong *et al.* [1] proposed a general framework to analyze the temporal structure of live broadcast sports videos. They formulated the structure analysis as the problem of detecting the fundamental views by using supervised learning and domain-specific rules. To generalize the semantic video indexing to unstructured multimedia data (e.g. entertainment, scientific, commercial, etc.), research efforts recently extended to develop generic models representing semantic concepts using advanced statistical pattern recognition and learning techniques. Milind R. Naphade *et a*l. [2] presented a probabilistic framework of multijects and multinets to map low-level features to high-level semantics and explicitly model the interconceptual relationships. C. Neti *et al*. [3] exploited multiple modalities (visual, audio, speech, text) to construct a multimedia semantic learning framework. They tried to train content depicting semantic concepts (events, objects, scenes) and make use of the resulting statistical models to classify unknown video content. Compared with structured video analysis using a domain specific model, such unstructured video analysis is more generic but exhibits lower efficiency.

As a semantic index, events are semantically meaningful and significant to users, whereas they are mostly specific to particular domains. S. Moncrieff *et al.* [4] tried to detect violent events and car chases in feature films by performing the analysis of environmental sounds such as gunfire, engines, horns, and explosions. Y. Rui *et al.* [5] developed effective techniques to detect excited announcers' speech and baseball hits from noisy audio signals, and fuse them to extract exciting segments of baseball programs. Y.P. Tan *et al.* [6] measured the variation and persistence of recovered camera motion to generate basketball video annotations, such as fast break, full court advances, etc. In [7], D.Q. Zhang *et al.* made use of caption text extraction and recognition to identify events such as ball count and game score in baseball videos. Since a content creator uses visual, auditory, and textural channels to convey meaning, multi-modality analysis is the future of semantic indexing.

In this paper, we propose a fusion scheme of visual and auditory modalities to detect events with strong semantic meaning in sports video. Instead of using Hidden Markov Models (HMM) based multi-modal integration methods [8][9], we introduce the semantic shot classification scheme to locate the segments with potential events, within which game specific sound, commentator speech, and environmental sounds are detected and recognized by a hierarchical Support Vector Machine (SVM) classifier. Subsequently, we heuristically associate camera shots' potential semantic linkage and sounds to infer predefined events occurrence. At present, we have successively identified five interesting events (i.e. serve, reserve, ace, return, score) in tennis videos. To fuse the visual and the auditory, we have to face two challenges: 1) video shots are not necessarily aligned with audio segments; 2) specific sound recognition is sensitive to noise and environmental sound. Practically, the proposed fusion scheme has provided a feasible solution in the sports video domain by: 1) heuristically aligning of semantically meaningful video shots and audio segment of interest, 2) avoiding unnecessary sound analysis in video shots unrelated to event detection.

The rest of this paper is organized as follows. In section 2, we choose tennis video as an example to describe the fusion scheme of visual and auditory modalities. In Section 3, we briefly introduce a unified framework for semantic shot classification in sports videos. A hierarchical SVM classifier for mid-level auditory representation is then discussed in Section 4. Experimental results are shown to evaluate the performance of event detection in tennis video by the multi-modality analysis in Section 5. Finally, this paper is concluded in Section 6.

2. A FUSION SCHEME OF VISUAL AND AUDITORY MODALITIES

We take the example of tennis video to explain this scheme. As listed in Table 1, we can summarize the tennis video shots in several major or interesting classes. With the help of production rules and game specific knowledge, it is straightforward to furnish each shot class with coarse semantics, such as play/break. To detect events with strong semantic meaning, performing in-depth analysis within a shot is necessary. Since autonomous segmentation and tracking are misleadingly difficult tasks in low-level computer vision, we take advantage of audio information to facilitate intra-shot analysis. Let us consider the following semantic sound transition patterns in court view shot of tennis video:

The sounds are summarized in four classes: hitting ball, silence, applause, and commentator speech. By computing the ball hitting times and the intervals between two ball hits, and checking the applause sound at the end of shots, we have come up with five basic events: serve, reserve, ace, return, and score (See [11]). Based on these events, we can recognize the five patterns in court view shot of tennis video as listed in Figure 1. Each pattern has strong semantic meanings.

Compared to soccer video, tennis video exhibits more canonical sound transition patterns, which intuitively conveys our perspective on the fusion of visual-based shot classification and auditory-based sound classification. An example of audio-visual fusion scheme for tennis video is illustrated in Figure 2.

3. A UNIFIED FRAMEWORK FOR SEMANTIC SHOT CLASSIFICATION

As shown in Figure 3, the proposed framework [10][12] makes use of domain knowledge of specific sport to perform a top-down video shot classification. According to the domain model, we try to learn the rules for shot classes identification. Rules have the great benefits that they are easily interpreted and can be used to choose the distinguishing features for training the classifiers. To determine the classifier, we can use various learning procedures and the following mid-level features fusion is executed at the shot level instead of key frame level.

As illustrated in Figure 3, this framework consists of four hierarchical levels. First, we derive low-level features (e.g. texture, motion vector fields, DC images) directly from compressed video data and fully decompress I-frames for further segmentation. Second, we exploit camera motion analysis, color-driven and texture driven spatial segmentation to produce a set of mid-level features such as camera motion pattern, dominant object motion, and homogeneous regions. Third, we analyze and reorganize collected mid-level features within each shot, thus mapping mid-level features to high-level semantic video shot attributes. Finally, we classify each shot into one of the predefined shot categories with learned classifier.

Compared with existing works, the proposed framework has the following unique features:

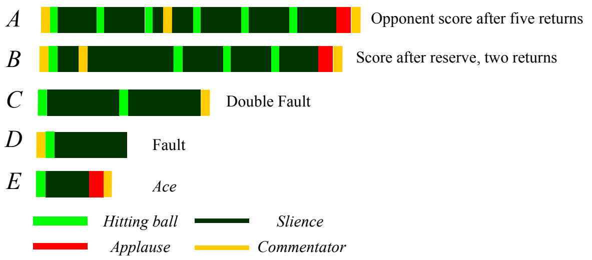

Figure 1: Five typical sound transition patterns within the court view shots in tennis video, along with semantic meanings

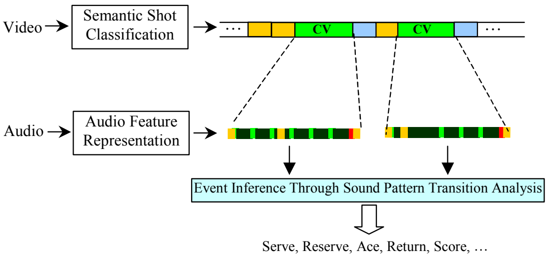
Video
Semantic Shot Classification
CV
CV
Audio
Audio Feature Representation
Event Inference Through Sound Pattern Transition Analysis
Serve, Reserve, Ace, Return, Score, …

Figure 2: An illustration of the fusion scheme for tennis video

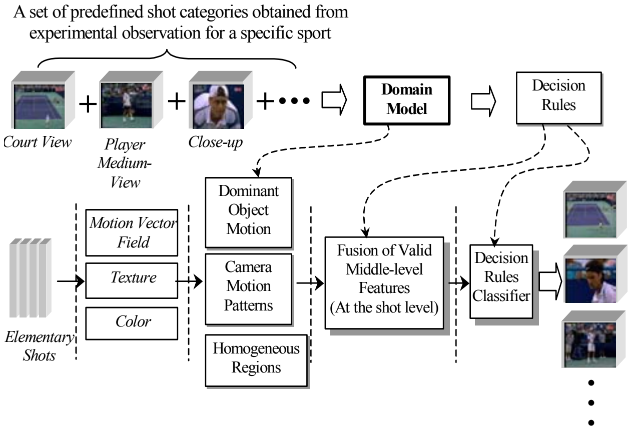
A set of predefined shot categories obtained from experimental observation for a specific sport
Court View
Player Medium-View
Close-up
Domain Model
Decision Rules
Elementary Shots
Motion Vector Field
Texture
Color
Dominant Object Motion
Camera Motion Patterns
Homogeneous Regions
Fusion of Valid Middle-level Features (At the shot level)
Decision Rules Classifier

Figure 3: A unified framework for sports video shots classification

- Direct relationship between shot classes and clear semantic meanings
- Adaptive mapping from low-level features to mid-level features
- Effective mid-level visual features representation according to domain model
- Fusion of valid mid-level visual features at the shot level under the supervision of decision rules.
- Real-time performance by the combination of pixel domain processing and transform domain processing.

The proposed framework has been tested over four types of sports videos: tennis, basketball, volleyball, and soccer. Good classification accuracy 85~95% has been achieved. For more details, readers are referred to [10]. To help understand the relationships between shot classes and potential semantic linkage, we give some statistics of shot classes in tennis videos in Table 1.

Table 1: Shot classes based on experimental observation for tennis

Type	Class	Ratio (%)	Major Class (Tick Besides)	Potential Semantic Linkage
Tennis	Close-up (CU)	37.5	√	Break, Serve
	Court View (CV)	29.8	√	Play
	Player Medium-View (PMV)	16.8	√	Break
	Audience (AV)	9.2		Active Reaction
	Bird-View (BV)	1.0		Begin a new game
	Replay (RP)	4.3		Exciting moment
	Undefined	1.4		

4. MID-LEVEL AUDIO FEATURES REPRESENTATION

4.1 Low-level Features Extraction

4.1.1. Mel frequency Cepstral Coefficients (MFCC)

The mel-frequency cepstrum has proven to be highly effective in automatic speech recognition and in modeling the subjective pitch and frequency content of audio signals. The mel-cepstral features can be illustrated by Mel-frequency Cepstral Coefficients (MFCCs), which are computed from the FFT power coefficients. A triangular band pass filter bank filters the power coefficients. The filter bank consists of K=19 triangular filters. They have a constant mel-frequency interval, and cover the frequency range of 0Hz – 20050 Hz. Figure 4 (a) shows the first four MFCCs of an audio segment from tennis video.

4.1.2. Zero Crossing Rate (ZCR)

In the context of discrete-time signals, a zero crossing is said to occur if successive samples have different algebraic signs. The rate at which zero crossings occur is a simple measure of the frequency content of a signal. This average zero-crossing rate gives a reasonable way to estimate the frequency of sine wave. Zero crossing rate is suitable for narrowband signals, but audio signals may include both narrowband and broadband components. Hence, we make use of zero crossings to distinguish between applause and commentator speech. Figure 4 (b) shows the ZCR of an audio segment from tennis video.

4.1.3. Others

Apart from the MFCC and ZCR, there are many other features for characterizing different kinds of music signals, such as linear prediction coefficient (LPC), short time energy (ST), spectral power (SP), and cepstral coefficients (CC) [13]. However, we did not use these features according to out test results (See Section. 4.2).

4.2 Feature Selection

Feature selection is important for discriminating audio signals. To select good features suitable for the classification of hitting ball, applause, silence, and commentator speech, we make use of a single-layer SVM classifier to evaluate the performance of a single feature in the classification. We summarize the results in Table 2. The audio signal samples were collected with 44.1kHz sample rate, stereo channels and 16 bits per sample from 5 tennis games with the total length of 15 minutes. The original signals were segmented at 20ms/frame, which is the basic unit for feature extraction. By using MFCC, we obtained satisfactory classification of hitting and silence. According to Table 2, however, it is difficult to achieve good classification of applause and commentator speech with one feature only. Hence, we consider isolating these two classes from other classes, and then further discriminating them from each other. Moreover, we have carried out experiments to evaluate the performance of different features in the classification of applause and commentator speech. According to the results in Table 3, we can see that MFCC and ZCR work well. Therefore, we choose MFCC and ZCR to design classifiers (See Section 4.3).

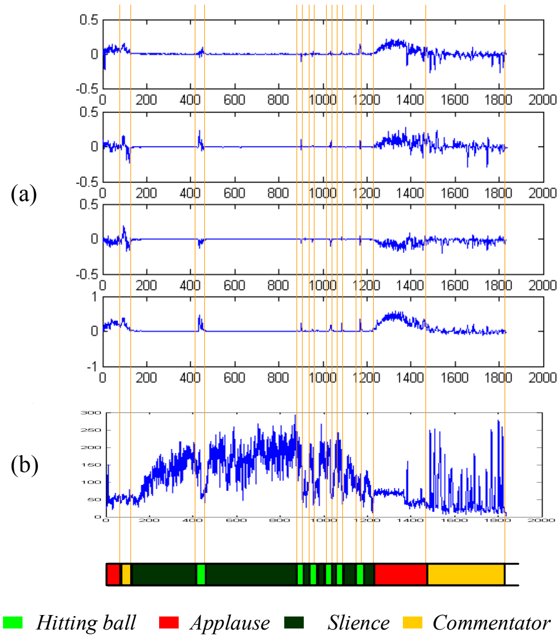

Figure 4: Low-level audio features within one court view shot in tennis video: (a) MFCC, (b) ZCR.

Table 2: Performance comparison (Misclassification Rate) of low-level feature in the classification

	Applause (%)	Commentator Speech (%)	Silence (%)	Hitting (%)
LPC	20.12	37.29	12.21	2.32
CC	19.23	21.89	12.39	2.44
MFCC	11.78	15.23	4.67	2.60
SP	17.47	26.88	15.88	2.72
ZCR	39.62	26.88	34.51	2.60
ST	66.19	33.35	71.89	2.60

Table 3: Performance comparison of low-level feature in the classification of commentator speech and applause

	MFCC	ZCR	SP	ST	CC	LPC
Misclassification Rate (%)	9.68	16.85	22.28	51.17	22.45	27.87

Table 4: Performance of the two-layer SVM classifier

	Applause	Commentator speech	Silence	Hitting
Error Rate (%)	7.23	11.29	7.62	1.1

4.3 A Hierarchical SVM Classifier

Based on the above analysis, we propose a two-layer hierarchical SVM classifier as shown in Figure 5. The kernel functions are $K(x,y)=\exp(-\|x-y\|^2/c), with\ c = 0.5$ · The performance evaluation is listed in Table 4. Note that sound itself is a continuous existence and humans normally make decisions about sound characteristics within a certain time chip. Hence, we exploit the sliding window technique to vote the sound type from a sequence of frame-based classification results (the ball hitting sound even covers 3~4 frames).

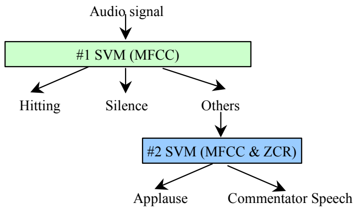

Figure 5: A hierarchical classifier for audio classification

5. EXPERIMENT RESULTS

According to the fusion scheme above, we have come up with a customized solution for detecting events. Table 5 lists the results from five video clips (each clip corresponds to a game). Note that these five events are not exclusive from each other. For example, a reserve itself is a serve; ace is a serve that the opponent cannot return. The events inference has considered context information. To help understand the relationships between events, we give the Figure 6. For more details, readers are referred to [11].

Table 5: Performance evaluation of event detection in tennis videos (5 games, 25 Minutes)

	Score	Reserve	Ace	Serve	Return
Ground Truth	49	34	19	84	131
No. of Miss	2	3	7	5	12
No. of False	6	3	0	1	8

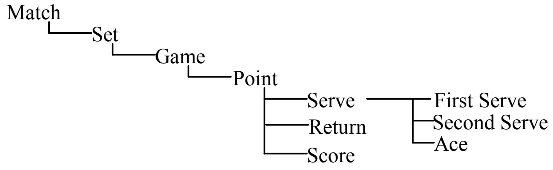

Figure 6: A hierarchical structure of tennis video

6. CONCLUSION

We have described an effective fusion scheme of visual and auditory modalities to detect events in sports video. Due to the canonical sound transition pattern and compact events, we choose tennis video as a test bed and have successfully identified five interesting events with strong semantic meanings by the proposed fusion scheme. Compared with existing multi-modality integration methods, this fusion scheme has provided a straightforward solution for event detection in structured videos, which features computational efficiency and operational flexibility. Furthermore, our experiment has shown that the audio signal analysis does facilitate the semantics mining through audio-visual compensation action. That is, we use visual cues to locate interesting video segments, and then use full-fledged audio processing techniques to bridge the semantic gap between available visual features and the richness of user semantics.

At present, we are extending this fusion scheme to other typical sports video (i.e. volleyball, basketball, and soccer), among which soccer video is the most challenging one. This comes from two challenges: 1) various camera angles and perspectives, 2) an extremely complex audio track. To some extent, the proposed shot classification framework has addressed the first challenge. As for the latter, we are planning to optimize the composition of low-level features and seek a set of generic mid-level audio features representation through machine learning techniques.

7. REFERENCE

[1] D. Zhong, S. –F. Chang, “Structure Analysis of Sports Video Using Domain Models”, In Proc. of *IEEE International Conference on Multimedia Expo*, 2001.

[2] M. R. Naphadem, and T. S. Huang, “A Probabilistic Framework for Semantic Video Indexing, Filtering, and Retrieval”, *IEEE Transactions on Multimedia* 3(1): 141-151, 2001.

[3] http://www.research.ibm.com/compsci/Iyengar_VisionDataTalk_external.ppt

[4] S. Moncrieff, C. Dorai, and S. Venkatesh, “Detecting Indexing Signs in Film Audio for Scene Interpretation”, In *Proc. of IEEE International Conference on Multimedia Expo*, 2001.

[5] Y. Rui, A. Gupta, A. Acero, “Automatically Extracing Highlights for TV Baseball Programs”, In *Proc. of ACM Multimedia*, pp. 105-115, 2000.

[6] Y.-P. Tan, D. D.Saur, S. R.Kulkarni, and P. J. Ramadge, “Rapid Estimation of Camera Motion from Compressed Video with Application to Video Annotation”, *IEEE Transactions on Circuits and Systems for Video Technology* 10(1): 133-146, 2000.

[7] D.Q. Zhang, R. Raj, S.-F Chang, “General and Domain-specific Techniques for Detecting and Recognizing Superimposed Text in Video”, In *Proc. of IEEE International Conference on Image Processing*, 2002.

[8] A.A. Alatan, A.N. Akansu, and W. Wolf, “Multi-modal Dialogue Scene Detecting Using Hidden Markov Models for Content-based Multimedia Indexing”, *Multimedia Tools and Applications* 14(2): 137-151, 2001.

[9] S. Eickeler and S.Muller, “Content-based Video Indexing of TV Broadcast News Using Hidden Markov Models”, In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 2997-3000, 1999.

[10] L.-Y. Duan, M. Xu, and Q. Tian, “Semantic Shot Classification in Sports Video”, to appear on *SPIE Storage and Retrieval for Media Database* 2003.

[11] http://www.hickoksports.com/glossary/gtennis.shtml

[12] L.-Y. Duan, M. Xu, X.-D. Yu, and Q. Tian, “A Unified Framework for Semantic Shot Classification in Sports Videos”, In *Proc. of the ACM Multimedia* 2002.

[13] N. C. Maddage, C. Xu, C.-H. LEE, and M. Kankanhalli, “Statistical Analysis of Musical Instrument”, In *Proc. of IEEE Pacific Rim Conference on Multimedia* 2002.