CREATING AUDIO KEYWORDS FOR EVENT DETECTION IN SOCCER VIDEO

Min Xu, Namunu C. Maddage, Changsheng Xu, Mohan Kankanhalli, Qi Tian

Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore, 119613
{xumin, maddage, xucs, tian}@i2r.a-star.edu.sg
School of Computing, National University of Singapore, Singapore, 117553
mohan@comp.nus.edu.sg

ABSTRACT

This paper presents a novel framework called audio keywords to assist event detection in soccer video. Audio keyword is a middle-level representation that can bridge the gap between low-level features and high-level semantics. Audio keywords are created from low-level audio features by using support vector machine learning. The created audio keywords can be used to detect semantic events in soccer video by applying a heuristic mapping. Experiments of audio keywords creation and event detection based on audio keywords have illustrated promising results. According to the experimental results, we believe that audio keyword is an effective representation that is able to achieve more intuitionistic result for event detection in sports video compared with the method of event detection directly based on low-level features.

1. INTRODUCTION

The rapid development of various affordable technologies for multimedia content capturing, data storage, high bandwidth/speed transmission and the multimedia compression standards such as JPEG and MPEG, have resulted in a rapid increase of the size of digital multimedia data collections and greatly increased the availability of multimedia contents to the general user. Sports video is one of such kind of multimedia data that attracts a global viewer-ship and research effort has been focused on semantic event detection in sports video to facilitate access and browsing.

Most of event detection methods in sports video are based on visual features [1-2]. However, audio is also a significant part of sports video. In fact, some audio information in sports video plays an important role in semantic event detection. Compare with research done on sports video analysis using visual information, very little work has been done on sports video analysis using audio information. Chang *et al.* [3] developed a speech analysis approach to detect football touchdowns. Keywords spotting and cheering detection were applied to locate the meaningful segment of video. Vision-based line-mark and goal-posts detection were used to verify the results obtained from audio analysis. Rui *et al.*, [4] proposed a solution to extract highlights from TV baseball program using audio-track features alone. To deal with an extremely complex audio track, speech endpoint detection technique in noisy environment was developed and support vector machines was applied to excited speech classification. A combination of generic sports features and baseball-specific features were used to detect the specific events. Zhang *et al.* [5] proposed an approach to detect the cheering event in a basketball video game using audio features. A hybrid method was employed to incorporate both spectral and temporal features. Sadlier [6] developed a method to summarize sports video using pure audio analysis. The audio amplitude was assumed to reflect the noise level exhibited by the commentator and was used as a basis for summarization.

The previous methods tried to detect events in sports video directly based on low-level features. Unlike complex low-level features, the mid-level representations are able to facilitate the high-level analysis from the semantic concept point of view. In this paper, we attempt to explore a novel framework using audio keywords to assist event detection in sports video. Some of audio mid-level representations are created from low-level audio features using supervised learning and used to detect semantic events in sports video by applying heuristic mapping. So we call these audio mid-level representations as audio keywords which are the reference points for event detection. We have applied this method to event detection in tennis video [7]. In this paper, we will use audio keywords to detect events in soccer video which is more complex than tennis video in structure.

2. AUDIO KEYWORDS IN SOCCER VIDEO

Keywords refer to some significant or descriptive words or some words used as a reference point for finding other words or information in text mining and analysis. We introduce the concept of keywords from text domain to sports video to represent middle-level features used for

0-7803-7965-9/03/$17.00 ©2003 IEEE

bridging the gap between low-level features and high-level semantics. We have successfully identified five typical sound transition patterns and combine them with semantic shot classification results to detect five interesting events in tennis video [7]. Unfortunately, soccer videos are much looser with less-canonical sound transition patterns for obviously and definitely telling all interesting events. It is hard to get determinate events by only checking the appearance of these audio keywords. However, an encouraging discovery is that, with the help of production rules and game specific knowledge, it is straightforward to furnish some potential semantic linkages from audio keywords to interesting events as shown in Table 1.

Table 1: Potential Linkage from audio keywords to events based on experimental observation for soccer

Audio Keywords	Potential Events
Long-whistling	Start of free kick, penalty kick, or corner kick Game start and end
Double-whistling	Foul
Multi-whistling	Referee reminding
Excited commentator speech	Goal or Shot
Excited audience sound	Exciting moments
Plain commentator speech and audience sound	Normal

To detect potential events, we create some audio keywords for whistling, commentator speech and audience sound because they are directly related to referees, commentators and audiences' actions in soccer games. Particularly, referees' whistling presents some certain judgments and instructions when the game is going. Therefore, we create seven audio keywords for soccer games: long-whistling, double-whistling, multi-whistling, exciting commentator speech, plain commentator speech, exciting audience sound and plain audience sound which have strong hints to events and are clearly shown in Figure 1. We create these audio keywords from low-level audio features using supervised learning method.

3. AUDIO KEYWORDS CREATION

Since audio keywords are created from low-level audio features, how to select proper features to best represent audio keywords is extremely important.

3.1 Low-level Features Extraction

In soccer video, the audio signal mainly comes from speech and sounds of commentator, audience, whistling, and environment. Therefore, we first extract some low-level features that are successfully used in speech analysis and then experiment whether they can provide good results for audio signal analysis in soccer video.

3.1.1. Zero Crossing Rate (ZCR)

In the context of discrete-time signals, a zero crossing is said to occur if successive sample 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 is suitable for narrowband signals, but audio signals may include both narrowband and broadband components.

3.1.2. Spectral Power (SP)

For a audio signal $s(n)$, each frame is weighted with a Hamming window *h(n)*, where N is the number of samples of each frame. The spectral power of the signal $s(n)$ is calculated as

$$S(k) = 10\log_{10}\left[\frac{1}{N}\left\|\sum_{n=0}^{N-1} s(n)h(n)\exp(-j2\pi\frac{nk}{N})\right\|^{2}\right] \qquad (1)$$

3.1.3. Linear Prediction Coefficient (LPC)

The basic idea behind linear predictive analysis is that auditory sample can be approximated as a linear combination of past samples. By minimizing the sum of the squared differences (over finite interval) between the actual auditory samples and the linear predictive ones, a unique set of predictor coefficients can be determined. The importance of linear prediction lies in the accuracy with which the basic model applies to vocal signals in sports audio.

3.1.4. LPC-derived Cepstral Coefficient (LPCC)

In the linear prediction analysis, the generative model is an all-pole filter with the transfer function $H(z) = 1/\sum_{i=0}^{p} a_i z^{-i}$, where p is the number of poles and $a_0 = 1$. The filter coefficients $\{a_i : i = 1,\ldots,p\}$ are chosen to minimize the mean square filter prediction error summed over the analysis window. From LPC, LPCC can be calculated by the following recursion $c_n = -a_n + \frac{1}{n}\sum_{i=1}^{n=1}(n-i)a_i c_{n-i}$, $n = 1,\ldots N$, where N is the number of cepstral coefficients. One of the advantages of LPCC is that it is decorrelated so that a diagonal covariance could be used in SVM to model statistical properties of the signals.

3.1.5. Mel-Frequency Cepstral Coefficient (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 19 triangular filters. They have a constant mel-frequency interval, and cover the frequency range of 0Hz – 20050 Hz.

3.2 Low-level Feature Selection

Feature selection is important for discriminating audio signals. To select suitable features for the classification of soccer audio keywords, we make use of a single-layer support vector machine (SVM) classifier [8] to evaluate the performance of a single feature in the classification. In our experiments, we use two-hour data of FIFA world cup 2002 which are equally divided for training and testing. We summarize the results in Table 2-4. Our selected features (ZCR, SP, MFCC, LPC, STE and LPCC) highlight both time domain and frequency domain properties of the audio signals. From Table 2, we find that whistling is easy distinguished by ZCR because the frequency of whistling is higher than other signals. Since the noise level of both commentator speech and audience's sound is high, individual feature cannot achieve good classification results for these two classes. By combining LPC and SP, the error rate is reduced to 14.15%.

Table 2: Performance comparison of low-level features in the classification of three main classes

Error Rate	LPC (%)	LPCC (%)	MFCC (%)	SP (%)	STE (%)	ZCR (%)
SVM1 (Whs vs Comm,Aud)	5.49	6.02	4.90	6.21	4.88	2.52
SVM2 (Comm vs Aud)	16.28	25.81	29.56	16.76	50.53	41.28

Table 3: Performance comparison of low-level features in the classification of subclasses of commentator speech (Exciting vs Plain)

	LPC	LPCC	MFCC	SP	STE	ZCR
Error Rate (%)	36.73	26.27	21.56	38.19	52.73	28.81

Table 4: Performance comparison of low-level features in the classification of subclasses of audience sound (Exciting vs Plain)

	LPC	LPCC	MFCC	SP	STE	ZCR
Error Rate (%)	41.31	37.92	26.68	24.85	51.34	22.57

3.3. Keywords Creation

We design a hierarchical SVM classifier for classification three main classes: Whistling, Commentator speech and Audience sound. For each main class, different methods and features are selected for subclasses identification. By computing the times of coterminous whistling and the duration of each whistling, we successfully detect three audio keywords of whistling. However, for the subclasses of commentator speech and audience sound, it is difficult to achieve good classification results using single feature. Hence, through feature comparison experiments, we use LPCC, MFCC and ZCR which work better than other low-level features for excited and plain commentator speech classification, and get 81.59% correct rate. Table 3 shows the results of individual feature for classification. Moreover, ZCR, MFCC and SP together provide good results of 79.83% correct rate in classification excited and plain audience sound. Performance of a single feature is shown in Table 4. Note that sound itself has a continuous existence and human normally makes decision about sound characteristics within a certain time segment. Hence, we exploit sliding window technique to vote the sound type from a sequence of frame-based classification results. Our framework of audio keywords creation is shown in Figure 1

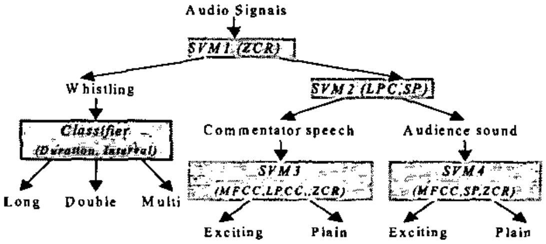

Figure 1: The framework of audio keywords creation

4. AUDIO KEYWORDS FOR EVENT DETECTION

In this section, we illustrate some cursory event detection by only using created audio keywords. Trying to find some corresponding sound combination and transition patterns for some interesting events, we get some promising results of detecting five events such as, free kick/penalty kick, foul, goal, shot, and game start/end in terms of rule-based heuristic mapping of keywords to high-level semantics. The framework we proposed is shown in Figure 2. Figure 3 shows the rules of events detection by audio keywords in soccer video.

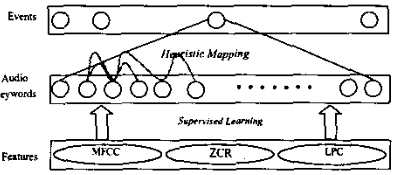

Figure2: The framework of event detection by audio keywords in soccer

RULE1: IF *The interval between double-whistling & long-whistling <= Threshold1*
Then Free kick or Penalty kick
RULE2: IF *Double-whistling*, **Then** Foul
RULE3: IF *Long-whistling & the time of exciting commentator speech and exciting audience sound alternate frequently>=Threshold2*
Then Goal
RULE4: IF *Excited commentator speech & exciting audience sound*
Then Wonderful shot or wonderful pass or wonderful save
RULE5: IF *Duration of Long-whistling>=Threshold3*
Then Game start or end
RULE6: IF *Double-whistling & announcer's speech*
THEN Substitution

Figure3: Heuristic decision rules for events detection by audio keywords in soccer video

According to these heuristic rules, we can detect some interesting events with strong semantic meaning in soccer video. Some results of event detection are shown in Table 6, Section 5.2.

5. EXPERIMENTS

5.1. Keywords Creation

We split audio stream from two hours soccer game which is selected from 3 matchers of FIFA World Cup 2002. The audio samples were collected with 44.1 kHz sample rate, stereo channels and 16 bits per sample. The audio signals were segmented at 20ms/frame, which is the basic unit for feature extraction. Based on observation, in most time, the sounds of some kinds of classes such as commentator speech and audience's sound are mixed together in soccer video. We label every sample by the class of which the sound occupies the dominant role. We used two third of these samples for training and one third for testing. The results of keywords creation are shown in Table 5.

5.2. Event Detection

We get satisfactory results for some important events by using heuristic decision rules listed in Figure 3. However, to distinguish free kick and penalty kick is almost impossible by only using audio keywords. In this paper we combine these two kinds of events into one class for detection. Some results are listed in Table 6.

Table 5: Performance evaluation of keywords creation

	LW	DW	MW	EC	CC	EA	CA
Error Rate (%)	7.27	10.89	8.93	23.58	20.24	26.37	24.15

LW: Long Whistling; DW: Double Whistling; MW: Multi-Whistling; EC: Exciting Commentator Speech; CC: Calm Commentator Speech; EA: Exciting Audience Sound; CA: Calm Audience Sound

Table 6: Performance evaluation of event detection in soccer videos (Three matches) A: Foul; B: Free kick/Penalty kick; C: Goal; D: Shot; E: Game start and end

	A	B	C	D	E
Ground Truth	129	11	12	86	12
No. of Misses	36	4	1	15	0
No. of False	12	1	3	19	2

5.3. Discussion

From Table 6, we can see that audio keywords work well for some given events detection. However, soccer videos have limited sound patterns that cannot cover all interesting events occurred in soccer games. Moreover, the event detection accuracy depends on the performance of audio keywords recognition. So only using audio keywords cannot completely finish the task of events detection in soccer video. To make event detection more accurate, we need to seek help from video analysis.

6. CONCLUSIONS AND FUTURE WORKS

We have presented an effective framework of creating audio keywords and using audio keywords to detect events in soccer video. Since soccer video features loose events, it is more difficult to identify sound transition patterns for event detection than other sports video such as tennis video. Comparing with other event detection methods, we provided a novel way to create audio keywords and further with the help of domain specific knowledge to detect interesting events.

At present, we are extending the work of audio keywords to visual keywords creation in soccer video and attempting to combine audio and visual keywords together to detect more interesting events in soccer video. In the future, this framework will be extended to other sports video.

6. REFERENCES

[1] C. Wu, Y.F. Ma, H.J. Zhang and Y.Z. Zhong, "Event Recognition by Semantic Inference for Sports Video," In *Proc. of IEEE International Conference on Multimedia and Expo*, Lausanne, Switzerland, pp. 805-808, Aug. 26-29, 2002.

[2] S.F. Chang, W. Chen and H. Sundaram, "Semantic Visual Templates: Linking Features to Semantics", In *Proc. of IEEE International Conference on Image Processing*, Chicago, IL, vol.3, pp. 531-535, Oct. 1998.

[3] Y.L. Chang, W. Zeng, I. Kamel and R. Alonso, "Integrated Image and Speech Analysis for Content-Based Video Indexing", In *Proc. of IEEE International Conference on Multimedia Systems and Computing*, pp. 306-313, 1996.

[4] Y. Rui, A. Gupta and A. Acero, "Automatically Extracting Highlights for TV Baseball Programs", In *Proc. of ACM Multimedia*, Los Angeles, CA, pp. 105-115, 2000.

[5] D. Zhang and D. Ellis, "Detecting Sound Events in Basketball Video Archive", Technical Report, Columbia University, https://www.ctr.columbia.edu/~dpwe/courses/e6820-2001-01/projects/dpzhang.pdf.

[6] D.A. Sadkier, S. Marlow, N. O'Connor and N. Murphy, "MPEG Audio Bitstream Processing towards the Automatic Generation of Sports Programme Summaries", In *Proc. of IEEE International Conference on Multimedia and Expo*, Lausanne, Switzerland, vol.2, pp. 77-80, Aug. 26-29, 2002.

[7] M. Xu, L. Duan, C. Xu and Q. Tian, "A Fusion Scheme of Visual and Auditory Modalities for Event Detection in Sports Video", Submitted to *IEEE International Conference on Acoustics, Speech, & Signal processing*, Hong Kong, China, April, 2003.

[8] V. Vapnik, *Statistical Learning Theory*, Wiley, 1998.