Automated Referee Whistle Sound Detection for Extraction of Highlights from Sports Video

P. Kathirvel, Dr. M. Sabarimalai Manikandan and Dr. K. P. Soman
Center for Computational Engineering and Networking
Amrita School of Engineering, Coimbatore Campus
Amrita Vishwa Vidyapeetham, India- 641105

ABSTRACT
This paper proposes a simple and automated referee whistle sound detection (RWSD) for sports highlights extraction and video summarization. The proposed method is based on preprocessor, linear phase bandpass finite impulse response (FIR) filter short-time energy estimator and decision logic. At the processing stage the discrete audio sequence is divided into non-overlapping blocks and then amplitude normalization is performed. Then, a bandpass filter is designed to accentuate referee whistle sound and suppress other audio events. Then, the filtered signal is fed to short-time energy (STE) estimator which includes amplitude squarer and linear filter to obtain a positive signal. In this work, we use decision rules based on the amplitude-dependent threshold and time-dependent threshold for detecting of referee whistle sound regions. The performance of the proposed design is tested using a large scale audio database including American football, soccer, and basket ball. The total duration of the test audio signal is approximately 12 hours and 11 minutes. The proposed method results in time-instants of boundaries of whistle sounds and then time instants are used to automatically extract the sports highlights from the unscripted video. Then, audio perception of the extracted sound segments is performed to indentify the false positive (FP) and false negative (FN). The proposed method has a detection failure rate of 19.4% (42 FP and 26 FN) and detects 324 whistle sounds successfully. The sensitivity and reliability of the proposed design are 92.5% and 80.5%, respectively. The design is implemented in MATLAB 7.0 version environment with the following system specifications: Intel (R) Pentium (R) Dual Quad CPU @ 2.40 GHz and 3 GB of RAM. The computation time is approximately 0.3-second for processing of 1-second block.

General Terms
Multimedia, Audio Classification, Content-based audio and video retrieval and summarization

Keywords
Audio classification, video summarization, sports highlight extraction, semantic video analysis, audio content analysis

1. INTRODUCTION
The design of sport highlights extraction system (SHES) is an emerging field of multimedia research that can be attempted with content-based audio and image analysis. Recent advances in digital computer technology, particularly in storage device technology, and video search engines have resulted in significant increases in the number and quality of multimedia databases such as movies, TV shows, comedy programs, ceremony videos, songs and sports [1]-[5]. These broadcast videos are the popular entertainment programs enjoyed by lots of people [2]-[7]. Generally, particular multimedia contents have been frequently searched and played or accessed from the digital multimedia libraries. In movies, for example, scenes like comedies, songs, and fights are extremely popular. The limited resources such as wireless channel capacity and time allocation, and human interest have created a strong demand for automatic extraction of highlights from the huge multimedia databases [1].

In recent years, sports video such as baseball, cricket, football, golf and soccer appeals to large audiences [2]-[5]. Generally, video consists of image sequences and audio tracks which provide visual and audio information of the program, respectively [2]. In sports video, audio stream generally includes announcer commentary (excited commentator speech and plain commentator speech), referee whistle sound, ball-hit, player speech, audience speech, music, crowd cheers and applause, and several environmental sounds [2]-[15]. Thus, the audio track of the sports video is a mixture of various sound sources. Consequently, sports videos usually have lots of background noise while movie videos and talk show videos have less background noise except some background musical sounds. The excited commentator speech, referee whistle sound, ball-hit, audience cheers and applause sounds are the typical events in sports video [2]. The audio keywords that are related with interesting events in a video are to extract highlights from a particular sports video [7]-[13]. Although sport highlights can be extracted using the information of image sequences and audio track, extraction of highlights is commonly done with the audio stream since that requires less computation and memory. Furthermore, in video summarization and searching applications, audio keywords of the audio stream of the video can be sufficient to extract highly semantic events present in unscripted videos.

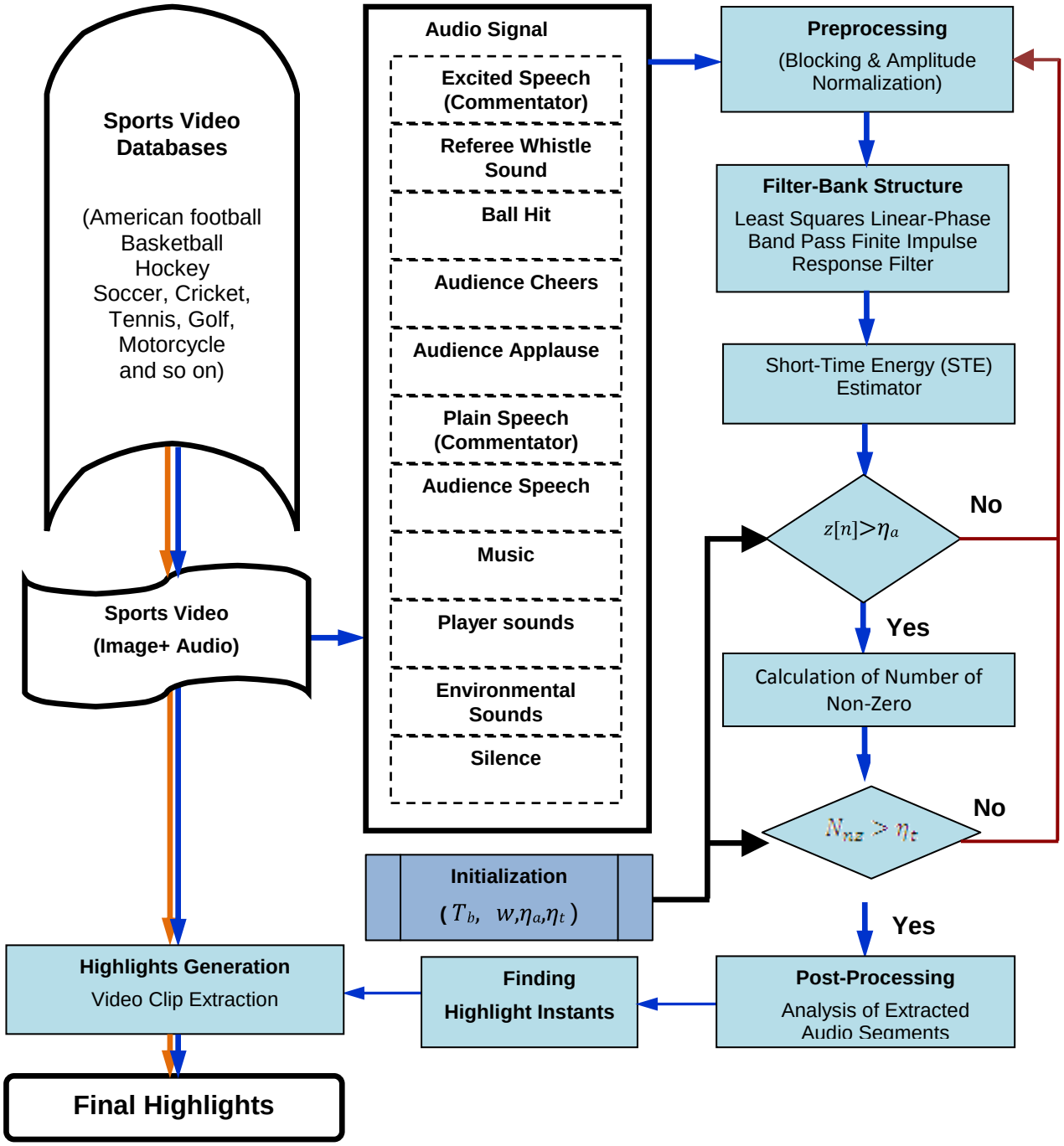

Figure 1. Overall Framework for Whistle Sound Detection

The observation shows that the referee whistle sounds, and the audience cheers and applause sounds are the most important audio keywords across different sports [13]-[15]. These sounds convey more information than the other sounds present in the audio stream. Therefore, in this work, whistle sound detection is presented and used to extract highlights from an unscripted sport video. Different methodologies have been employed for detection of important audio events from a continuous audio stream [1]-[5]. However, a performance of audio classification is impaired by audio mixtures and different kinds of background noise. This paper describes a filter-bank based referee whistle sound detection approach. The rest of the paper is organized as follows. Section 2 describes the proposed whistle sound detection method. Section 3

presents the experimental results and the performance of the propose method. Finally, conclusions are drawn in Section 5.

2. MATERIALS and METHODS

To extract highlight segments from the full sports video effectively and rapidly, the RWSD is designed based on the preprocessing, filter-bank, short-time energy estimation, amplitude-dependent threshold and time-dependent threshold comparisons, and sound instants determination. The overall framework for whistle sound detection is shown in Fig.1. The different stages and the predefined parameters are described in the following subsections.

2.1 Preprocessing

Because of the limited system resources and time varying characteristics of the audio signals, preprocessing is most important to develop an efficient RWSD approach. As mentioned in the introduction section, the video is a compound of image sequences and audio tracks. Since processing of the image sequences commonly requires more memory space and computation time, the video is divided into visual and audio channels and then audio signals present in the single-channel is processed in this approach. The duration of the audio signals is usually longer. The memory space required for storing the audio signals depends on the sampling rate and bit resolution. In this approach sampling rate of 16000 samples per second and amplitude resolution of 16 bit per sample are used for digitization of the audio signals by considering the system requirements such as memory space and computational load. The discrete audio sequence is divided into non-overlapping blocks. The duration of block is chosen based on the mean duration of referee whistle sounds that usually present in sports video. Here, the duration of the block (*T_b*) is 6 seconds, and each block is processed separately for audio feature extraction.

The mean removal and amplitude normalization is performed at the preprocessing stage and can be implemented as

$$x_i[n] = z_i[n] - \mu_i, \quad i = 0,1,2,3,.....L-1$$
$$n = 0,1,2,3,.....N-1$$

$$y_i[n] = \frac{x_i[n]}{\max(|x_i[n]|}$$

where L denotes the number of blocks, N denotes the number of samples in each block, μ_i is the mean value of the i^{th} block, and the $x_i[n]$ and $y_i[n]$ denote the zero-mean original and normalized discrete audio sequences, respectively for the i^{th} block. The normalized discrete sequence $y_i[n]$ is fed to filter-bank which accentuates the desired sound components and suppresses other background noises.

2.2 Band Pass Filtering

The normalized mixed audio sequence is processed using a least squares linear-phase bandpass finite response (FIR) filter. Several different choices of bandpass filters for audio signal processing have been described for analyzing various sounds. A high pitched sound such as a referee whistle is very different from other sounds such as audience applause and commentator speech (excited and plain). Therefore, detection of the audio events is performed using the filter-bank structure which generally consists of more than two filters. The bandpass filters are designed according to the desired spectral characteristics. The filter-bank structure of an audio event-based highlight extraction system includes a least square linear-phase bandpass FIR filter for emphasizing the referee whistle sound components. The linear-phase FIR filter is designed based on the square error criterion. This technique is straight-forward and is applicable to arbitrary desired frequency responses that minimizes the weighted, integrated squared error between an ideal piecewise linear function and the magnitude response of the filter over a set of desired frequency bands. Since the technique is simple, non-iterative and optimal with respect to square error criterion, this design philosophy is adopted in this work to derive the filter coefficients.

The cut-off frequencies of the bandpass filter are chosen based on the spectral characteristics of the whistling devices. From the spectral analysis of various audio signals, lower and upper frequencies of the dominant spectral components of the referee whistle sound is measured. Then, these frequencies are used for designing the bandpass FIR filter to extract referee whistle sound from a mixture of different sound sources such as audience cheers or applauses, speech and music. The frequency range for the whistle sound detection is approximately is 3750 Hz- 4100 Hz. This is determined empirically to give good results over a variety of referee whistle sounds. The signal obtained at the output of the bandpass filter is a compound of desired sound source and sound components from other sound sources that lie in the desired spectrum. However, magnitude of the desired audio signal is significantly larger as compared to the background noises. Therefore, the bandpass filter output is utilized to derive the feature signal which can be used for detecting particular audio event. Since the output of the filter can be a bipolar signal, memoryless nonlinear transformation is applied to the filter output followed by linear filtering to provide a unipolar signal in the next stage.

2.3 Short-Time Energy (STE) Estimation

In order to detect the referee whistle sound segments, short-time energy (STE) is considered as a basic audio feature in this design. A simple short-time energy estimate is performed using nonlinear squaring and linear filtering operations. The short-time estimator is computed as

$$s[n] = \sum_{k=n-N+1}^{n} f^2[k]\, h[n-k]$$

where the filtered signal $f[n]$ is fed to amplitude squarer and linear filter with a finite rectangular impulse response $h[k]$ and

produces a feature signal $s[n]$. The digital linear filtering provides a necessary smoothing to the squared signal with large response for portions corresponding to the desired sound. The smoothed STE signal is searched for local maxima and will be used to detect the desired sound segments at the decision stage. The smoother behavior of $s[n]$ can be studied by varying the window size (W) to reduce ripples and multiple peaks before detection. The choice of window size results in tradeoff between false-positives and false-negatives. Large window size provides better detection accuracy but computation load is high. Since the detection of desired sound segments is the important task, small window size is used in this design. Note that this approach may not be useful to determine instants of whistle sound exactly but thresholding rules employed at the decision stage detects desired sound events successfully. If the end-points of the detected sound events are needed then audio segments are further processed with larger window size. Then, smoothed feature signal is input to the instant marker. This process may reduce the computational load since the number of whistle sounds is usually small. It is well known that a high recognition accuracy of the desired events leads to good highlight extraction.

2.4 Decision Rule

Amplitude-dependent threshold and time-dependent threshold comparisons are used to detect the whistle sound events at the decision stage. Firstly, amplitude-dependent threshold comparison is done, where energy values of the feature signal $s[n]$ are compared with the user-specified amplitude-threshold (η_a). Amplitude-dependent thresholding rule sets any energy value less than or equal to the threshold to zero. The output of the amplitude-dependent thresholding stage cannot be used to derive decision logic since the STE estimator output may have short-duration large spikes corresponding to undesired audio events that have overlapping spectrum. It can be observed that duration of the referee whistle sound is normally 0.5 to 3.5 seconds. The detector decision based on the output of the amplitude-dependent threshold leads to more false-positives, and thus detection failure rate will be high in this case. In order to increase detection accuracy, the time-dependent threshold is applied on the output of amplitude-dependent thresholding stage. For implementation of second thresholding rule, number of non-zero values N_{nz} in the thresholded sequence is calculated. Then, the resultant N_{nz} value for each block is compared with the user-specified time-dependent threshold η_a. The audio segments whose N_{nz} values are above threshold η_t are selected as final highlight features for scene recognition and segmentation. It can be seen that the performance of the detector design depends on selection of two thresholds that are employed at the decision stage. However, optimal threshold values can be determined if the characteristics of the desired audio events are well-known. Thus, study on whistle sound parameters such as intensity and duration are most important to design a better decision logic for recognition of desired sound in a mixture of various audio events especially in the single-channel case.

3. RESULTS and DICUSSION

In this application note, referee whistle sound detector is designed for exploring the possibility to build a unified framework to extract highlights from sports videos. We can observe that the referee whistle sound, the excited commentator speech, the audience cheers and applauses are common events and more general across various sports. Moreover, human usually pay more attention to these scenes and thus audio events relating to interesting segments in spots videos are detected for highlights generation. In this note, we focus on the referee whistle sound as the audio cue for highlight extraction work. Therefore, a simple referee whistle sound (RWS) detector is presented in the previous section. The detection accuracy of the RWS detector is studied for the American football videos. The experimental database comprises sixty referee whistle sounds extracted from football, American football and soccer, and full audio tracks of American football video. The total duration of the audio tracks is approximately 2 hours and 11 minutes. To evaluate the performance of detector, benchmark parameters are used including false negative (FN) which means failing to detect a true audio event (actual referee whistle sound), and false positive (FP) which represents a false sound detection. True positive detection (TP) stands for correct recognition of sound present in the input mixed signal extracted from the audio track of the sport video. The FP detection represents a detector error of desired sound identification that doesn't exist in the analyzed signal and the FN detection represents a detector error of missed whistle sound that exists in the analyzed signal. By using FN, FP and TP, the sensitivity (Se), positive predictivity (+P), detection error rate (DER) and reliability (Re) are calculated using the following equations, respectively.

$$Se = \frac{TP}{TP + FN} \times 100$$

$$+P = \frac{TP}{TP + FP} \times 100$$

$$DER = \frac{FP + FN}{TS} \times 100$$

$$\mathrm{Re} = \frac{TS - (FP + FN)}{TS} \times 100$$

where *TS* denotes the total number of desired sounds in the test data. The detection results for the proposed design are summarized in Table I. We test the proposed sound detection on the continuous audio stream of a 2-hour and 11-minute American football game. The American football video has high background noise from the audience cheers and applauses, excited commentary and various environmental sounds. The audio signals are all single-channel, 16 bit per sample with a sampling rate (*Fs*) of 16 kHz. In the case of whistle sound event detection, audio sequence is divided into non-overlapping blocks, and each 6-sec

Table 1. Referee whistle sound detection results for audio stream extracted from American football game

Detector Specifications			Total Sounds (TS)	True Positives (TP)	False Positives (FP)	False Negatives (FN)	DER (%)	Se (%)	+P (%)	Re (%)
Window size (W)	Amplitude Threshold(η_a)	Time Threshold (η_t)								
0.3*FR	10	0.2*FR	126	98	23	28	40.48	77.78	80.99	59.52
0.3*FR	10	0.3*FR		89	14	37	40.48	70.63	86.41	59.52
0.3*FR	10	0.2*FR	103#	94	27	09	34.95	91.26	77.69	65.05
0.3*FR	**10**	**0.3*FR**		**89**	**14**	**14**	**27.18**	**86.41**	**86.41**	**72.82**

Note: # denotes the total number of referee whistle sounds after eliminating some of sounds that are characterized as single beat with low-intensity in the annotation file, and the quality ratings of those sounds are less than 3.

block is processed for detection of referee whistle sound. The thresholds used at the decision stage are found empirically. After detecting the referee whistle sound event, timing of the event is stored and then these timing instants are used in the scene segmentation stage.

Experiment shows that the timing instants corresponding to whistle sound events often deviate from the actual locations of the desired events. Therefore, we include a certain number of seconds of video clips before the beginning and ending moment of the desired event to generate final highlights. Finally these segments are compared to those ground-truth highlights that are labeled by human viewers. To compare the timing instants computed for each whistle sound the referee whistle sound in audio tracks of American football is annotated manually. For the following detector specifications *W=0.3*Fs*, amplitude-dependent threshold value of 10, and time-dependent threshold value of *0.3*Fs*, timing instants of the extracted sounds, false positives and false negatives are given for the tested 1310 blocks with 6-sec in duration. After removing some whistle sounds with low intensity and poor quality in audio perception, timing instants of total referee whistle sounds of 103 are marked. False positive and false negative are identified by audio perception of the extracted sound segments. The proposed design has a detection failure rate of 27.18% (14 FP and 14 FN). The sensitivity and positive predictivity of this design are 86.41% and 86.41%, respectively. The proposed design detects 89 referee whistle sounds correctly and produces 14 false positives due to the excited commentator speech present in the test blocks. If the time-dependent threshold value is *0.2*Fs*, the sensitivity of the design is better but it has more false positives. The number of false positives can be further reduced by using similarity measure at the post-processing stage. For the above optimal design specifications, the proposed method is tested using a large scale audio database with duration of 12 hours and 11 minutes, which includes American football, Soccer, and Basket ball. The overall performance of the proposed method is shown in Table 2. The proposed method achieves a sensitivity of 92.5%, a positive productivity of 82.6%. The method has failure detection rate of 19.4% which includes the 42 false positive detections and the 26 false negative detections. Experiments show that the proposed method is more suitable for sports highlights extraction and video summarization applications.

Table 2. Performance of the proposed method

Database (12-hr and 11-min)	TS	TP	FP	FN	Se (%)	+P (%)	Re (%)
American football, basket ball, Soccer	350	324	42	26	92.5	82.6	80.5

4. CONCLUSIONS

This paper presents a simple and automated referee whistle sound detection based on the preprocessor, linear phase band pass finite impulse response (FIR) filter short-time energy estimator and decision logic. The performance of the proposed methodology is validated using a large scale audio database including American football, soccer, and basket ball. The total duration of the test audio signal is approximately 12 hours and 11 minutes. The proposed method results in time-instants of boundaries of whistle sounds and then time instants are used to automatically extract the sports highlights from the unscripted video. Then, audio perception of the extracted sound segments is performed to indentify the false positive (FP) and false negative (FN). The proposed method has a detection failure rate of 19.4% (42 FP and 26 FN) and detects 324 whistle sounds successfully. The sensitivity and reliability of the proposed design are 92.5% and 80.5%, respectively. The design is implemented in MATLAB 7.0 version environment with the following system specifications: Intel (R) Pentium (R) Dual Quad CPU @ 2.40 GHz and 3 GB of RAM. The computation time is approximately 0.3-second for processing of 1-second block. As a continuation of this work, we are collecting audio signals of different sports videos and studying spectral characteristics of the referee whistle sounds present across different sports. Furthermore, we are designing different band pass filters for detection of more general audio events like audience cheers and applauses, ball-hit, and excited commentator speech to provide a better extraction of highlights from a sport video.

5. ACKNOWLEDGMENT

The authors would like to thank the Editor-in-Chief and the anonymous referees for their valuable suggestions and comments.

6. REFERENCES

[1]. C. Xu, J. Wang, H. Lu, and Y. Zhang, "A novel framework for semantic annotation and personalized retrieval of sports video," *IEEE Trans. on Multimedia*, vol. 10, No. 3, pp. 421-436, April 2008.

[2]. Z. Xiong, R. Radhakrishnan, A. Divakaran, and T.S. Huang, "Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, pp. 632 – 635, 2003.

[3]. P. Chang, M. Han, and Y. Gong, "Extract highlights from baseball game video with hidden Markov models," in *Proceedings of International Conference on Image Processing (ICIP)*, vol. 1, pp. 609–612, 2002.

[4]. Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," *Eighth ACM International Conference on Multimedia*, pp. 105 –115, 2000.

[5]. 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.1026–1039, May 2006.

[6]. A. Hanjalic, "Generic approach to highlight detection in a sport video," in *Proceedings of IEEE International Conference on Image Processing (ICIP)*, vol. 1, pp. 1–4, 2003.

[7]. X. F. Tong, H. Q. Lu, Q. S. Liu, and H. L. Jin, "Replay detection in broadcasting sports video," in *Proceedings of 3rd International Conference on Image and Graphics*, pp. 337-340, 2004.

[8]. I.Otsuka, R. Radharkishnan, M. Siracusa, A. Divakaran, and H. Mishima, "An enhanced video summarization system using audio features for a personal video recorder," *IEEE Trans. Consumer Electron.*, vol. 52, no. 1, pp. 168–172, Feb. 2006.

[9]. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Trans. Image Processing*, vol. 12, no. 7, 2003.

[10]. N. Babaguchi, Y. Kawai, and T. Kitahasgi, "Event based indexing of broadcasted sports video by intermodal collaboration," *IEEE Trans. Multimedia*, vol. 4, no. 1, pp. 68–75, 2002.

[11]. H. Pan, B. Li, and M. Sezan, "Automatic detection of replay segments in broadcast sports programs by detection of logos in scene transitions," in *Proc. IEEE ICASSP*, 2002.

[12]. D. Zhang and S. F. Chang, "Event detection in baseball video using superimposed caption recognition," in *Proc. ACM Multimedia*, pp. 315–318.

[13]. R. Cai, L. Lu, H.-J. Zhang, and L.-H. Cai, "Highlight sound effects detection in audio stream," in *Proc. IEEE ICM.*, 2003, vol. 3, pp. 37–40.

[14]. R. Jarina, J. Olajec, "Discriminative feature selection for applause sounds detection," in Proc. *8th Int. Workshop on Image Analysis for Multimedia Interactive Service* , Greece, 6–8 June 2007, pp. 13–16.

[15]. M. Xu, L. Duan, C. Xu, M. Kankanhalli, and Q. Tian, "Event detection in basketball video using multi-modalities," in *Proc. IEEE Pacific Rim Conf. Multimedia*, Singapore, Dec. 15–18, vol. 3, pp. 1526–1530, 2003.