SSD'14 1569846987

The Importance of Audio Descriptors in Automatic Soccer Highlights Generation

Arnau Raventós, Raúl Quijada, Luis Torres, Francesc Tarrés
Signal Theory and Communications Department
UPC - Barcelona Tech. Castelldefels (Barcelona)
Email: francesc.tarres@upc.edu

Eusebio Carasusán, Daniel Giribet
Televisió de Catalunya
Esplugues de Llobregat (Barcelona)
Email: ecarasusan.q@tv3.cat

***Abstract*—Automatic generation of sports highlights from recorded audiovisual content has been object of great interest in recent years. The problem is indeed important in the production of second and third division leagues highlights videos where the quantity of raw material is significant and does not contain manual annotations. Many approaches are mostly based on the analysis of the video and disregard the important information provided by the audio track. In this paper, a new approach that combines audio and video descriptors for automatic soccer highlights generation is proposed. The approach is based on the segmentation of the video contents into shots that are further analyzed in order to determine its relevance and interest. These video-shots are scored taking into account the fusion between different audio and video features. The paper is mainly focused to emphasize the importance of audio detectors that play a key role in the analysis and scoring of the video-shots. Specifically, a new algorithm for referee's whistle detection is proposed. The algorithm has been proven to be very robust and efficiently discriminates professional whistles against other types of noises such as public cheering-up, music instruments, etc. Several results have been produced using real soccer video sequences that prove the validity of the proposed audio and video fusion scheme.**

***Index Terms*—video highlights, content analysis, audio descriptors, whistle detector, semantic detection, multimodal processing and fusion**

I. Introduction

Sport video highlights consist in creating a short video that retains the most interesting parts of the original match while disregards those events that may be considered of low interest by the viewer. Generation of automatic highlights is a key application that may reduce cost and time for many TV stations and Internet content providers that need to generate summaries of sport events just a few minutes after the end of the match. In some cases, the sequences are annotated and summarized manually by journalists but this operation represents a laborious and exhausting task that is only feasible for very special events such as finals or top matches. In practice, recorded sport events are continuously produced and many of them correspond to second and third divisions leagues where the highlights have to be generated from the raw audio and video tracks, without using any manual annotations. Therefore, there is a need for developing automatic sports highlights generation approaches and research in this area has been very extensively pursued during the last decade.

Due to its extraordinary popularity in many countries, soccer is one of the most significant sports where video highlight generation is being applied. A variety of approaches have been presented in the literature. One strategy followed in [1] and [2] is to find those video-shots containing low-level graphics, text, TV logos, etc., which may indicate that an event of interest has occurred and that the shot should be inserted in the highlights. However, these approaches are not useful in the professional market since most TV broadcasters and Internet content providers are subscribed to clean-feed video signals that do not contain production graphics nor commentator channels. Another approach is to find a set of low-level video descriptors that are applied to the video-shots of the original sequence in order to detect some typical frames such close-ups, medium shots, field lines, goal posts, etc. These descriptors are applied to an expert system that selects those shots that will finally be inserted into the highlights. The expert system may be based on empirical rules such as in [3] or trained with statistical classifiers as in [4] and [5]. In all of these approaches most of the analysis is done in the video domain and only very elementary audio descriptors, such as power-level, are sometimes taken into account. In fact, audio analysis has been proven to be very effective for sport summarization. A strategy based exclusively in audio features is proposed in [6] for selecting interest sport events in a personal video recorder.

In this paper, a multimodal approach that combines several audio and video descriptors is proposed. The general processing scheme includes all the elementary MPEG-7 audio descriptors plus more advanced detectors for tracking the evolution of the audio level inside a video-shot or between adjacent shots. Moreover, a novel and very efficient whistle detector is also proposed. The whistle detector is a very important indicator of the interest of an action of the game. Referees usually use the whistle in situations like faults, kick-off goals, etc.

This paper is mainly devoted to the audio analysis part of the proposed framework with special emphasis in the algorithm for detecting and discriminating the referee's whistle against other noises. It is divided in the following sections. Section II presents an overview of the overall system. The audio

This work has been partially supported by the Spanish administration agency CDTI under the project CENIT-Buscamedia (CEN-20091026) and the CICYT project TEC2011-22989 'PBInt'

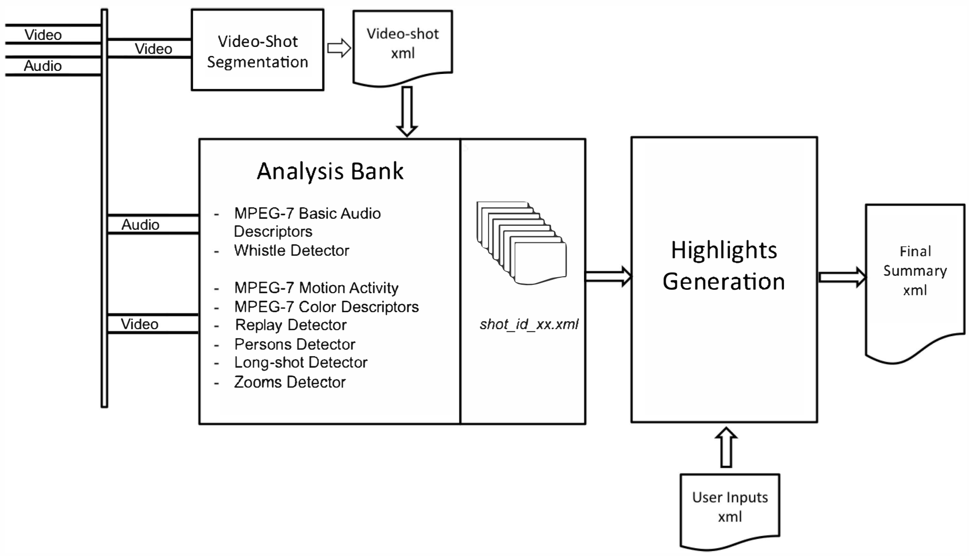

Fig. 1. General scheme of the analysis and highlights soccer generator framework

descriptors and detectors are explained in Section III. The main part of this section is devoted to the whistle detector algorithm. Section IV presents some results and performance of the system and finally, Section V contains the conclusions and future work.

II. Framework for Soccer Highlights Generation

Video-shots are defined as a series of continuous frames captured by a single camera that runs for a period of time. The whole video contents may be viewed as a sequence of non-overlapping shots. Soccer video highlights are made-up of a sequence of shots that collect the essential events of interest, usually in a linear timely basis where different shots are presented in the same order that have occurred. It is also assumed that the highlights are generated from the clean-feed that was produced during the live broadcasting of the game, containing the video and the stereo audio tracks without the commentaries channel. Clean feed sequences consist in live recordings that are contributed by the event producer to subscribed broadcasters. It includes the original sound captured by the microphones in the stadium and do not contain advertisement nor graphics. Broadcasters adapt this video sequence to their audience inserting commentators channels, graphics and advertisement. In clean-feed sequences the quality of the audio recording is not masked by the commentators channels and possibly the analysis of public noise and whistle is more reliable. Moreover, no auxiliary cameras, player-follow shots or alternative views are supposed to be available to generate the highlights. In this work, it is supposed that the minimum unit for building up the highlights will be the video-shot and that the whole clip will be a concatenation of selected video-shots found in the original sequence. The general diagram of the automatic soccer highlights generator proposed in this paper is represented in Fig. 1.

The Analysis Bank processes the audio and video tracks of every shot and determines a set of audio and video descriptors that will be annotated in XML format and transferred to the Highlights Generation module. This module processes the descriptors of each shot using a set of filters that try to detect different semantic situations, which may be considered as events of interest. The final score of a shot may take into account not only its own descriptors but also the ones of its neighbor's shots. The user may interact with the system specifying the total duration and the percentage of every semantic filter in the final summary.

A key component in the architecture of the system is the video-shot segmentation module that analyzes the video track and provides the shots time codes to the analysis bank. The approach presented here consists in employing two types of shot boundary detectors, one for hard cuts and the other for cross dissolves. The first detector consists in a classic histogram frame-by-frame comparison using the Chi-Square distance. The cross-dissolve boundary detector has been proposed by W.Abd-Almageed in [7] and is based on the rank analysis of a matrix composed by n-histogram frames through a Singular Value Decomposition (SVD) factorization. The hard-cut detector is used during the action of the game while the cross-dissolve is used before the beginning, in the half time and after the game. This strategy tries to exploit the style of production that is used nowadays in most soccer broadcasting. The type of boundary detector is automatically selected taking into account the results of the whistle detector that determines the beginning and end of each half. These shot boundary detection algorithms have been applied to different soccer video sequences. In particular, for a groundtruth of 3 soccer matches a recall of 95.2% and a precision of 98.8% for the abrupt transition detector, and a recall of 91.5% and a precision of 84.4% for the cross-dissolve detector have been obtained. These results show that the selected algorithms are adequate to be used in the overall scheme.

The analysis bank includes a variety of analysis tools that extract the audio and video descriptors. These tools have been optimized for real time execution in a medium-cost laptop computer. The video descriptors include three MPEG-7 color descriptors, the MPEG-7 Motion Activity descriptor, a persons detector, a replay detector, a long-shot detector and zoom detectors. Some of these descriptors are widely used in literature and their implementation have been based on available open-source code solutions. In other cases, such as in the person detector, the proposed solution uses a combination of several methods (a collection of adaboost detectors: face, profile, head&shoulders plus skin based detectors) and the final algorithm is tuned for best performance in this specific type of scenario (tilted views, non-frontal faces, etc.). The audio descriptors and detectors will be discussed in detail in Section III.

The Highlights Generation Module (see Fig. 1) is composed of multiple filtering processes that attempt to assign a semantic meaning to each shot. The shots are filtered using empirical knowledge rules and a final score is given to each shot.The *shot selection* stage produces the soccer highlights video sequences taking into account user's preferences such as the length of the summary and the percentage of appearance of selected semantic events. As an example, the user may specify a video highlights of 3 minutes where 50% of the time is dedicated to goals, 30% to important moments, 10% to goal-celebrations and 10% to faults. The *shot selection* algorithm gathers the shots with the highest scores for each filter, sorts them and finally selects a number of shots till the desired duration is achieved. When there are too few shots for a specific semantic event the system uniformly distributes the missing duration within the remaining events. An example of the scoring procedure and the type of filters is provided in Section IV.

III. Audio Descriptors and Detectors

The analysis bank includes some low-level and mid-level audio descriptors such as the MPEG-7 basic audio descriptors, a collection of audio power increase detectors (intra-shots and inter-shots) and an advanced whistle detector. All these descriptors are computed using a single audio track that is obtained as the mean value of the left and right channels using a sampling frequency of 48 kHz and 16 bits per sample.

A. MPEG-7 low level audio descriptors

The standard MPEG-7 defines a set of low-level audio descriptors with the purpose of facilitating interoperation among content-based retrieval applications. These descriptors are primarily designed for indexing audio media using probabilistic sound models. Audio descriptors may play a significant role in soccer highlights generation because many interest events are highly correlated with specific audio features such as an increase of volume or changes in frequency. An overview of MPEG-7 low-level audio descriptors may be found in [8]. The standardized descriptors include several measures of waveform, power, spectral characteristics, etc.; and different open-source implementations may be found. In this work, the MPEG-7 Audio Encoder library [9] has been used. This library was originally developed at the Institute of Communications Engineering at Aachen University and extended at the Università Politecnica delle Marche-Ancona. It includes a total of 22 low-level descriptors and 2 description schemes. The descriptors are computed on an audio frame time-basis using a hop rate of 10 ms. Therefore, for every video frame, a total of 4 samples of the audio descriptors are available. The analysis bank may be configured for computing only a reduced subset of the audio descriptors.

Generally, the descriptors that are computed are the *AudioWaveForm, AudioPower, AudioFundamentalFrequency, AudioSpectrumEnvelope* and AudioSpectrumCentroid. These descriptors provide a frame based scalar information about time and frequency characteristics. It is possible to detect significant changes in volume or in noise spectral characteristics following the evolution of these reduced set of descriptors. However, if desired, the user may specify any audio descriptor of the MPEG-7 standard implemented in the library.

B. Inter & Intra shot-based audio detectors

As mentioned before, the final soccer video highlights comprise a selection of video-shots that are considered as events of interest. However, the information collected by the MPEG-7 audio descriptors is sampled at a constant hop rate of 10 ms which is not appropriate to evaluate the relevance of the audio track in a complete video-shot. Therefore, it is necessary to post-process and filter these low-level descriptors in order to obtain a higher-level representation. The approach followed in this paper is that of detecting some special circumstances in the evolution of the audio descriptors within a video-shot or between adjacent video-shots. As an example, three detectors that measure the evolution of the audio power are described:

- The *A.Power.H* and *A.Power.VH* detectors represent peak levels of audio power within a shot, where H stands for high and VH for very high. Logical binary values are associated to those shots whose maximum audio power is over 95% and 97%, respectively, in reference to the maximum audio power value of the entire audio soccer track.
- *A.IntraInc.50* and *A.IntraInc.100* represent audio power increments within a shot. The low-level audio power descriptors are averaged for every second. The first detector is used to represent logical true values when the audio power in these averaged intervals is increased in 50% whereas the second detector represents increments of 100%.
- *A.InterInc.50* and *A.InterInc.100* is the same as A.IntraInc but they refer to average audio power increments between contiguous shots.

The same strategy used for detecting special situations in the *AudioPower* descriptor may be implemented for other descriptors such as the *AudioFundamentalFrequency*, the *AudioSpectrumCentroid*, etc. These detectors may be able to find situations where the public is singing, cheering-up a team or booing a referee's decision. However, up to now, these frequency detectors have not yet been considered in the final video-shot scoring, and their results are not reported in this paper.

C. Whistle detector

This section presents the methodology used to detect the sound produced by a referee's whistle and to discriminate it from other sounds and noises. In many sports, such as soccer, the detection of the referee's whistle provides highly valuable information to detect events of interest. Whistle detection has been considered extensively in literature and different strategies have been proposed [10], [11]. Most of these proposals are based on analyzing the spectral content of the signal and detecting maximum energy in the whistle's frequency band. The analysis is usually performed in the frequency domain either by calculating the DFT transform at the band of interest or through the use of a filter bank. Although these methods produce acceptable results in some scenarios, the number of false detections may be significant in recordings with too much cheering noise nearby whistle's frequency band. The method proposed in this paper tries to improve these results by further exploiting the spectral characteristics of professional whistles.

The proposed strategy for whistle detection is sketched in Fig. 2 and is composed of 5 blocks. The function of every block is briefly outlined below.

The first block segments the audio signal into frames. In order to have enough frequency resolution the frames have a length of 100 ms.

The frequency analysis block computes a set of Discrete Fourier Transform (DFT) samples in the band of interest for every audio frame. That band is selected between 3.5 kHz and 4.5 kHz that broadly includes the frequencies produced by professional whistles. The reduced set of samples of the DFT is computed using the Goertzel algorithm [12].

The next stage estimates the energy contained in the interest band of an audio frame using:

$$E(m) = \sum_{k=K_1}^{K_2} |X_m(k)|^2 \tag{1}$$

where K_1 and K_2 represent the DFT samples at the limits of the interest region and m represents the frame time index.

Some approaches threshold the above energy estimate and use it as a simple whistle detector. Although there exists high correlation between energy peaks in Eq. 1 and whistles, in some difficult scenarios thresholding may produce an unacceptable number of false alarms due to the presence of noise in the band of interest. In fact, the overall energy in the interest band may increase due to several reasons such as supporter's cheering, singing, stadium public address, etc., which makes energy thresholding a low reliable method. Due to this problem, the last two stages in the whistle detector system try to reduce the number of false detections by analyzing the frequency contents of the candidate frames with more detail, looking not only to the total energy but also to the specific contents inside the interest band.

To discriminate frames containing a whistle from other high-energy frames, the spectral characteristics of the whistle have to be taken into account. Nowadays, the whistles used in most sports are known as pea-less whistles. The sound is produced when blowing into a cavity that contains 3 chambers that vibrate at slightly different frequencies. However, these frequencies are not stable and may vary slightly with the specific excitation, the direction of blowing and the whistle model. In order to measure the tonal structure of the spectrum in the interest band and to discriminate between tonal and wideband contents, an entropy-based stage is introduced. Wideband contents are usually produced by supporters singing or cheering, by ambient music, noise, etc. To discriminate the tonal vs. non-tonal nature of the audio frame, a kind of parallelism with the concept of entropy is applied. The method consists of considering a normalized version of spectrum samples in the interest band as being samples of a probability density function. When the entropy is computed on these samples, a number that indicates the spread of the samples in the frequency band will be obtained. High numbers indicate wide spread while small numbers indicate that the energy is concentrated on a few, high probable energy samples. This entropy-based concept is defined as:

$$H(m) = \sum_{k=K_1}^{K_2} \rho_m(k) \cdot log_2(\rho_m(k)) \tag{2}$$

$$and \;\; \rho_m(k) = \frac{|X_m(k)|^2}{\sum_{r=K_1}^{K_2} |X_m(r)|^2} \tag{3}$$

Applying a threshold to this entropy value permits to discard all those audio frames where the energy is spread over the interest band. The tonal audio frames that do not exceed that threshold will be further processed by the final stage.

The final stage consists of a functional analysis of the total number of peaks that exceed a threshold. Its purpose is to discard some sounds that are sometimes confused with the whistle. A typical example are whistles of the spectators who usually also have maximum energy in the same frequency band and also have a tonal characteristic. However, the difference is that human whistling only produces a single tone and not the 3-tone spectrum of a professional pea-less referee's whistle. The implemented algorithm is based on selecting all peaks that exceed a threshold and eliminating those not separated by at least 150 Hz. Only audio frames with 2 or 3 peaks are finally validated.

IV. Performance Evaluation

This section is divided in two parts. The first part is centered on the performance analysis of the whistle detector while the second reports some significant results that have been obtained in the automatic soccer highlights generation using a fusion of audio and video descriptors.

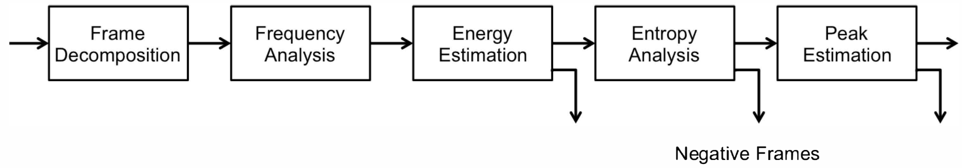

Fig. 2. Block diagram of the whistle detector algorithm

A. Whistle detector testing

Whistle detector performance has been evaluated using a database that contains a set of 5 complete recordings of soccer games. This database has been manually annotated registering those time-codes where the referee's whistle is present. Groundtruth annotation is performed using both video and audio tracks. Video track gives a helpful context required to discern referee's whistles among other sounds such as vuvuzelas, horns, supporter's whistles, etc. Aside from this database, a test signal has also been generated. This test signal is made up of a selection of some referee's whistles, already annotated in the database, in conjunction with other especially difficult sounds, usually inducing false positives, which have been encountered in these recordings. The total length of the test signal is around 60s and contains 10 referee's whistles, which are represented in Fig. 3.

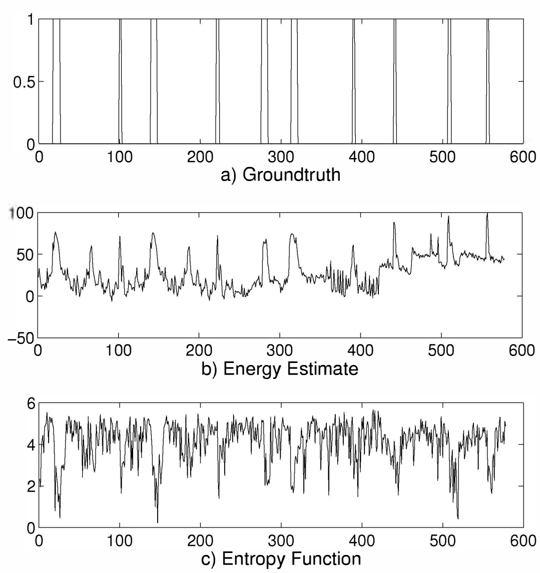

Fig. 3. Summary of results obtained by the different modules of the whistle detector for the test signal. a) Groundtruth b) Energy in the interest band c) Entropy estimation of the spectrum.

The energy estimation in the interest band, obtained in the third stage of the algorithm, is represented in Fig 3.b. This energy curve is used in some approaches, such as [10], [11], to detect the presence of the whistle when the signal exceeds a threshold. Comparing Fig. 3.a (whistle groundtruth) and Fig. 3.b a high correlation between whistle presence and energy peaks is observed. However, in some difficult scenarios, thresholding may produce an unacceptable number of false alarms due to the presence of noise in the band of interest. To evaluate the performance of this energy thresholding strategy we have chosen a threshold value to produce a *Recall* of 0.9 for the whole database and tested if the *Precision* obtained was satisfactory. Regrettably, the *Precision* was slightly below 0.6, which means that a significant number of positives are still false detections. Further analysis on misclassified samples show that the problem is produced by audio frames containing significant energy in the interest band due to the presence of other type of audio components (cheering, horns, music, etc.).

The results obtained when applying the complete proposed algorithm to the same database are notably improved. The entropy and peak estimation blocks permit to increase the *Recall* up to 93% maintaining a *Precision* of 88%.

B. Soccer highlights generation performance

The framework presented in Fig. 1 has been used to generate automatic highlights of 2 soccer matches. Apart from the *A.Power.H, A.Power.VH, A.IntraInc.50, A.IntraInc.100, A.InterInc.50* and A.*InterInc.100* detectors, which have already been described in Sec. III.B, other detectors have been used. These detectors include a binary detector of the presence of a referee's whistle. This detector has been applied in the current, previous and nex temporal video-shot. The system also includes three video descriptors that detect the presence of close-up shots of persons, the detection of zoom-ins and the detection of replays. All these detectors are weighted using an empirical model and give a final score to each video shot in the original recording. Those shots with the highest score are inserted in the final highlights up to the duration specified by the user.

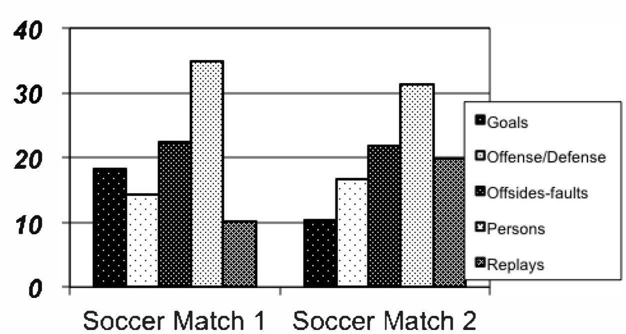

Fig. 4. Shots classification in 2 different soccer matches.

Fig. 4 shows performance of the filter's bank, where the soccer shots that constitute a summary are classified according to five groups: goals, offense/defense, offside-faults, persons and replays. The goals field contains the shots where a goal or a goal opportunity happens; opportunity of goal is any offensive play that ends close to the goal or due to a goalkeeper intervention. Offensive/defense represent the offensive and defensive actions that do not entail a goal or goal opportunity. Offside/faults groups the faults and offside. Persons are aided to represent the shots that contain public or persons as soccer players, coaches and other team members. Replays are shots that contains replays of faults, offside, goals and other interesting events. The shot classification of the soccer match 1 and 2 from the Fig. 4 highlights the shots distribution in five groups, where persons is the most popular shot in contrast to goals and replays that depend on the soccer match activity as can be observed from soccer match 1 that contains more goals and goals opportunities than soccer match 2 These results confirm that the scoring is quite efficient and that many interesting shots in the match are satisfactorily collected in the summary. Moreover, different filters may be defined using different detectors and different weightings. Detectors and weightings may be adapted to detect special situations such as goals, corners, etc.

From a subjective point of view the highlights are generally quite satisfactory and efficiently collect many interesting video-shots. For highlights durations of around 3-5 minutes the number of shots with low interest for the viewer are not significant. However, it may increase when the specified duration is about 30 minutes. The system is able to automatically include a significant number of goals in the highlights. In simulations with a total of 5 complete matches, 75% of the goals were included in the highlights, what is considered a good result by the journalists in charge of the summarization and annotation of the original soccer game.

V. Conclusions

This paper reports a framework for generating automatic video highlights of soccer video sequences. The method is based on segmenting the video sequence into shots and scoring them through a set of low and mid level audiovisual descriptors. The paper is focused on the importance of audio descriptors in the scoring of video-shots and weighting the interest of an action of the game. Basic MPEG-7 audio descriptors provide useful information about audio power, audio fundamental frequencies, and waveform. Mid-level detectors try to find significant changes in these basic descriptors in order to find interest video-shots. Experience has shown that audio power level significantly increases when an interest action occurs. The highest peaks in audio power may be used to detect goals. The evolution of the audio between adjacent shots is also extremely important to give a measure of the importance of the events.

The paper defines a set of mid-level audio detectors that are used, together with other video clues, to infer semantic contents and score the significance of video-shots. A robust whistle detector is also proposed. The algorithm uses a cascade of methods to verify if candidate frames may contain a whistle. The final performance is considered excellent and is effectively used to discriminate between referee's whistles and other non-professional whistles.

It also discriminates noise, cheering-up music, etc. Moreover, the whistle detector has been used to detect the beginning and ending of each part of the game allowing the use of a different shot-boundary detector adapted to the production of the different phases of the game. The proposed framework has been tested in a specific scenario and satisfactory results have been achieved.

Acknowledgments

The authors specially thank research assistants Pere Joaquim Mindán, Olga Mur and graduate student Albert Martín for their contribution to the design of the system and its software implementation. The soccer video sequences used for testing and evaluating the system have been kindly provided by Televisió de Catalunya (TVC).

References

[1] E. Fendri, H. Ben-Abdallah, and A. Ben Hamadou, "A novel approach for soccer video summarization," Second International Conference on Multimedia and Information Technology (MMIT 2010), pp. 138-141, April 24 - 25, 2010, Kaifeng, China.

[2] M.Y. Eldib, B.S. Abou Zaid, H.M. Zawbaa, M. El-Zahar and M. El-Saban, "Soccer video summarization using enhanced logo detection," in Proc International Conference of Image Processing ICIP 2009, pp 4345-4348, Nov 2009.

[3] A. Ekin, M. A. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," IEEE Transactions on Image Processing, vol. 12, no. 7, pp. 796-807, July 2003.

[4] H. M. Zawbaa, N. El-Bendary, A. Ella Hassanien, and T. Kim, "Event detection based approach for soccer video summarization using machine learning," International Journal of Multimedia and Ubiquitous Engineering, vol. 7, no. 2, pp 63-80, April, 2012.

[5] D. A Sadlier and N. E. O'Connor, "Event detection in field sports video using audio-visual features and a support vector machine," IEEE Transactions on Circuits and Systems for Video Technology, vol. 15, no. 10, pp. 1225-1233, Oct. 2005.

[6] I Otsuka, T. Radhakrishnan, M. Siracusa, A. Divakaran, and H. Mishima, "An enhanced video summarization system using audio features for a personal video recorder," IEEE Trans. on Consumer Electronics, vol 52, pp 168-172, Feb 2006.

[7] W. Abd-Almageed, "Online, simultaneous shot boundary detection and key frame extraction for sports videos using rank tracing", 15th IEEE International Conference on Image Processing, ICIP 2008, San Diego, California, USA, pp.3200-3203, October 12-15, 2008.

[8] H.G. Kim, N. Moreau, T Sikora, "Low-Level Descriptors", in MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval, 1st edition, New York: Wiley, 2005, ch.2, pp. 13-59.

[9] http://mpeg7audioenc.sourceforge.net

[10] Karthirvel, P; Sabarimalai, M.; Soman, P.; "Automated Referee Whistle Detection for Extraction of Highlights from Sports Video". International Journal of Computer Applications. Vol 12 – No. 11, January 2011.

[11] Mellinger, D.K.; Martin, S.W.; Morrissey, R.P.; Thomas, L.; Yosco, J.J.; "A method for detecting whistles, moans and other frequency contour sounds". Journal Acoustic Soc. Am. 129, June 2011.

[12] Oppenheim, Alan V.; Schafer, Ronald W. "Discrete-Time Signal Processing". Third Edition. Prentice Hall Signal Processing Series. 2010