Audio based Soccer Game Summarization

Helenca Duxans, Xavier Anguera and David Conejero
Telefnica Investigacin y Desarrollo
{hdb,xanguera,dco}@tid.es

***Abstract*—This papers presents a simple and near real-time performance system for detecting highlighted events of soccer game retransmissions and generating their video summaries. The proposed detection algorithm is based on two acoustic features of the audio track: the block energy and the acoustic repetition index. To the authors' knowledge, the acoustic repetition index has not been used previously in similar applications. This index represents the correlation between a narrow acoustic section and the seconds just after and before it, in order to detect sections of audio where repetitions occur. The system has been validated on a corpus with UEFA EURO competition games, achieving good scores in goal recall.**

I. Introduction

In this paper we present a new technique to detect goal and highlighted events in broadcasted soccer games based on acoustic features and a procedure to generate video summaries of the game. Hot spot detection in sport retransmissions has been studied from the video-only perspective, the acoustic-only perspective and also combining both media. The system presented in this paper relays on non-previously used acoustic features to perform the analysis, resulting in a simple (in contrast to other studies that required the training of a set of acoustic models) and an effective way to detect hot spots in soccer games.

There are multiple industrial applications of soccer game analysis, such as automatic summarization services for TV broadcasters, Internet users or instantaneously notification of hot spot events to specific mobile or Internet users. The system presented in this paper has been designed to cover two applications with two different targets, which will state the requirements of the system. On the one hand, an application which target are final users consists in the automatic generation, just at the end of the match, of a short summary of the event. Therefore, the near-real time performance is a requirement. On the other hand, an application which target are professional users performs a pre-selection of relevant moments (hot spots) in order to facilitate the manual generation of soccer game summaries by sport editors. Language independence and 100% of goal recall in medium duration summaries are the main requirements of this application.

The outline of this paper is as follows. In section II related work on highlighted moment detection in sports videos is reviewed. Then, in section III the design requirements of the current system are presented and the available material for the developement is explined in section IV. Section V reviews the study of the proposed acoustic features and in section VI the final design is presented. Finally, conclusions and future work are explored in section VIII.

April, 2009

II. Related Work

Highlight detection in sports videos is a well studied topic within the area of content-based video indexing (CBVI) [12]. It aims at automatically annotating and indexing video material for later retrieval or summarization of contents [10]. Given its clear commercial applications, extensive bibliography is found that studies the detection of events in sports. In this section those related to soccer highlights detection, which is the focus of this paper, will be analyzed.

In order to detect highlighted events in a soccer game one can differentiate between those publications studying it from an audio-only analysis perspective [8], [4], [15], [7], [2], from a video-only analysis perspective [1], [11] or, more generally, as a multimodal problem [13], [9], [3], [5], [16], [14], [6].

Multimodal systems can be further classified into those using one modality in a detector stage and the other modality as a refinement stage, in order to bring down false alarm rates [13], [3], [16], those fusing all modalities at the same level [5], [9] and those using multimodal features to obtain semantic features, used then to take decisions on the current event [14], [6]. In [3] pitch and energy are first used to find excited-audio segments, which are then separated from commercials by a posterior video analysis where evaluation of the amount of green color present in the shots is used. In [13] they also use pitch as a first step, and compute the ratio of shot changes and the presence/absence of goal-mouth as necessary for a goal-related highlight to occur. On the contrary, in [16] they first use image analysis to partition the video into shots, which later are mined to find goals using audio-visual features and decision trees.

Papers fusing all modalities include [9], which uses audio, video, textual features and static (a priori known) information and fuse them via a maximum entropy famework. In [5] a hierarchical structure is used, which the authors aim that reduces computational costs and avoids the use of shot boundary detection. In [6] they first convert from low level features to semantic features using multiple transforms and then use a two-dependence Bayesian network to fuse all information into a highlight detector. Similarly, in [14] they convert audio-visual features to keywords using SVM classifiers and then use a rule-based system to detect highlights. In general, whenever image processing is involved in the analysis it adds a level of complexity and computational cost to the system that might avoid it from running in realtime, although normally achieve levels of performance higher than using only one modality.

Competition	Season	Games	Commentators	Language	Source	Labels
UEFA EURO	08	15	2	Spanish	Cuatro	Available
Premier League	07-08	14	2	Spanish	La2 & Teledeporte	Not available
LFP	07-08	5	0	-	TV carrier	Not available
LFP	07-08	10	aprox. 3	Spanish	LaSexta & Telecinco	Not available
LFP	08-09	3	1	Catalan	TV3	Not available

TABLE I
CORPUS OF BOADCASTED SOCCER GAMES

In [11] they use only video information and study the patterns of camera movements in order to determine the kind of plays and whether there are highlights in the recordings. Similarly, in [1] the authors use camera motion only and an HMM classifier for the same task.

The algorithm presented in this paper uses audio-only descriptors to determine highlights in soccer footage. Similar research in the bibliography that use audio-only descriptors can be classified according to the audio nature they evaluate, whether they take into account the commentator level of excitement [7], [8] or just the public response [2], [15]. Such classification has a practical connotation in that broadcasted material usually includes commentator speech overlaid, where as initial recordings just include the public response.

In [8] several interesting features are extracted from the audio signal for highlights detection, namely energy related, phoneme-level, information complexity and prosodic features, whose combination and classification decision is done via SVM modeling. In [15] MPEG-7 features are used to train HMM-like models to classify highlights for soccer, baseball and golf events. Similarly, in [4] the authors compare the MPEG-7 Audio Spectrum Projection (ASP) features with MFCC features for detecting soccer goal events using HMM models. In [2] MFCC and HMM are paired again to classify soccer acoustic events using 6 pre-trained classes. All these systems need a priori labeled database to train classification models, with similar acoustic conditions for training like will be found in the test data. This is many times not robust as commentators might change and fields being recorded are many times different and with very different cheering crowds and acoustic characteristics. Finally, in [7] the authors use the Scale Factors in the MPEG audio bitstream to detect highlights based on the level of commentator or spectator excitement.

III. DESIGN REQUIREMENTS OF THE SYSTEM

The main goal of the study presented in this paper is to develop a system to detect highlighted events of soccer game TV retransmissions, with special interest on detecting the goal events, to generate summaries of configurable durations.

The focus of the study are TV broadcasted soccer games with at least one professional speaker as a commentator, without any restriction about the language of the retransmissions. Although the detection of the hot spot events must be performed in real time, the system will select the most relevant events once the entire game has been aired.

The design of the detector system must cover two main applications, both related to soccer game summaries generation but with different targets. The application whose target are final users consist in automatically generating soccer game summaries, as a technology enabler for automatic services of multimedia contents. Examples of such services are: MMS/e-mail soccer alerts, instantaneous soccer games summaries on Internet, interactive TV, etc.. Their main requisites are to include only relevant events and to generate very short summaries. The application whose target are professional editors consists in selecting highlighted events to be shown to professional editors in order to speed up the manual creation of soccer summaries. In that case, non relevant events are acceptable in the pre-selection while all the goal events must be present.

Based on the goal, the focus and the applications to be covered, the design requirements of the system are:

- Both the commentator and the ambient noise may be used.
- Language independent features.
- Near real-time computational response.
- Few training material and manual system adaptations.

In order to fulfill all the requirements we have designed an audio-based detector which operates with fast extraction language independent features.

IV. MULTIMEDIA MATERIAL

In order to determine the relevant features for hot spot detection in soccer games we have collected a broad corpus of TV retransmissions. Table I summarizes the contents of the corpus.

In a first instance, both retransmissions with and without commentators were inspected, in order to determine the main similarities and differences between them. Four of the five competitions were recordings of Spanish TV channels, five with national cover in Spanish (La2, Teledeporte, Cuatro, La Sexta and Telecinco) and one with regional cover in Catalan (TV3). The other competition consists in recordings distributed by a TV carrier without any commentators.

Only the material of the UEFA EURO has available labels to mark the time and type of highlighted events, extracted from a soccer specialized web page (http://www.eurocopa.com/) and manually checked to augment the time precision to 1 second. The available hot spots were limited to: goals, penalties, shots to goal, red cards, yellow cards and injuries. Any remarkable play not included in this classification, such as a conflictive off-side or a splendid dribbling, is not labeled as a highlight. Table II summarizes the available labels.

Although all the corpus has been inspected to look for relevant audio features and to subjective evaluate the soccer

Label	Apparitions
Goal	41
Penalty	2
Red card	2
Yellow card	54
Injury	2
Shot	216

TABLE II
LABEL DISTRIBUTION FOR THE UEFA EURO COMPETITION

game summaries, only the competition with labels allows to present numerical results of recall and precision figures.

V. ACOUSTIC FEATURES SELECTION

The key point in the design of the proposed algorithm was to find out those acoustic features which were easy computable (from the point of view of computational time and load) and also enough discriminative to allow the detection of hot spots in near real-time and with a high degree of goal recall and highlight precision.

Acoustic based systems can be designed according to two strategies: systems relaying on probabilistic models of acoustic features and systems relaying on scores related directly to acoustic features, without any probabilistic model assumption. Systems based on models usually have more requirements in terms of training material, since to estimate a general model requires training data captured in all the acoustic environments to be covered. However, we wanted to investigate if a simple algorithm, with low requirements, would be able suitable to the problem.

In order to find out which acoustic features will be useful to the hot spot detection task, the corpus presented in section IV was inspected. From audio inspection of the corpus we concluded that:

- The instant energy seems to play an important role in hot spots.

- The more "chaotic" the audio is, the probability that a hot spot occurs is higher.

- Without using any language dependent cue, the way the commentators speak is related to what is happening in the game.

In the following paragraphs the acoustic features related to each one of the three observations will be presented and further analyzed.

A. Block energy

In order to capture the audio energy evolution of the retransmissions the audio track is demultiplexed and divided into segments of one second length, with a 75% of overlap. The contents outside of the frequency range of $[660Hz-4.4KHz]$ are filtered out for each of the segments, and the energy of the resulting signal is computed.

The required resolution for the detector system was estimated to be about 1 second, since for some hot spots it is difficult to determine the exact instant of the event. Therefore, the obtained energy vector is decimated by four selecting the maximum value of the four adjacent points. Finally, the decimated values are processed by a median filter.

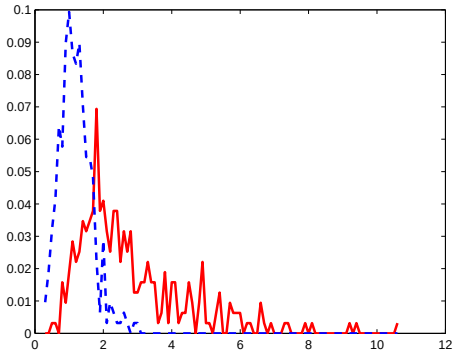

Fig. 1. Energy histogram: blue dashed line for neutral events and red continuous line for highlight events.

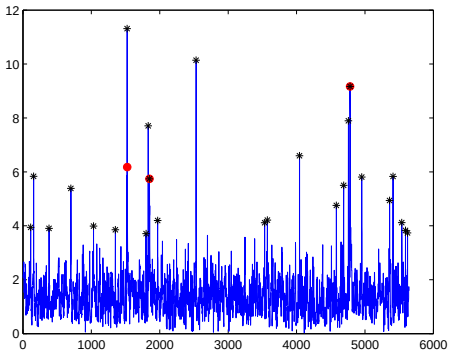

Fig. 2. Energy value evolution in a single retransmission. Black crosses mark the 25 highest points and red circles the real goals

Figure 1 shows the probability distribution of the energy values according to their classification into non-relevant or relevant points for the UEFA EURO subcorpus. Although both distribution share part of the energy space, it is clearly shown that the energy value is a good candidate to discriminate between hot spots and neutral moments of a soccer retransmission.

Moreover, the proposed system does not intend to be a classifier of sport events, but a detector of the most interesting moments in a single sport retransmission. Therefore, what is really interesting is that the energy values distribution inside a single game were less overlapped (see figure 2).

B. Entropy of power frequency distribution

The entropy of the power frequency distribution has been studied as an index of the chaos perceived in the audio track when a hot spot is found during the broadcast. The hypothesis is that when the power of the audio was distributed with a specific shape indicates that "a controlled" audio is perceived, but when the distribution is similar in all the frequency ranges a chaotic sound is perceived.

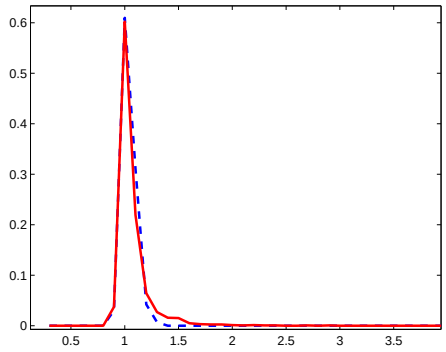

Fig. 3. Entropy histogram: blue dashed line for neutral events and red continuous line for highlight events.

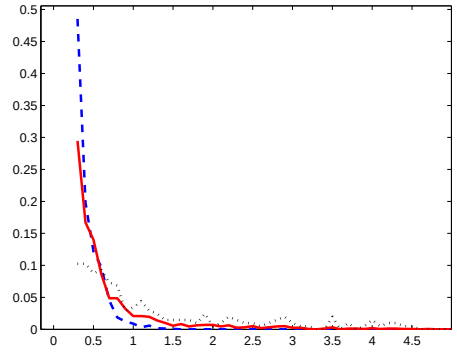

Fig. 5. Repetition index histogram: blue dashed line for neutral events, red continuous line for highlighted events and black doted line for goals.

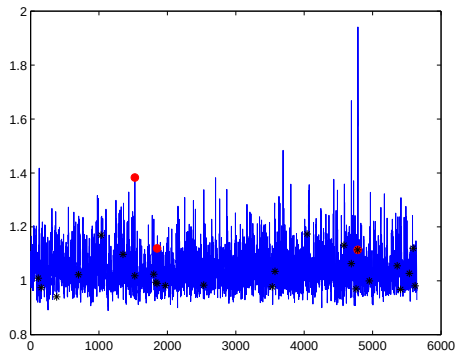

Fig. 4. Entropy value evolution in a single retransmission. Black crosses mark the 25 highest points and red circles the real goals

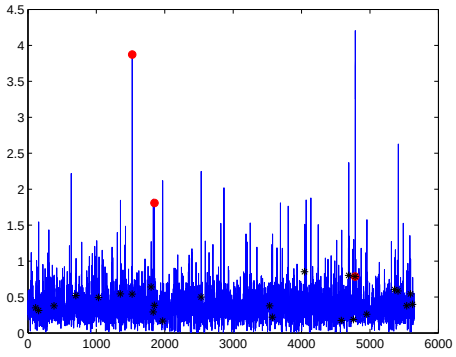

Fig. 6. Repetition index evolution in a single retransmission. Black crosses mark the 25 highest points and red circles the real goals.

Similar to the energy study, the entropy distribution between highlighted points and neutral points have been compared (see figure 3). The distributions do not corroborate the hypothesis, since both distributions are overlapped.

Moreover, the entropy values inside a single retransmission do not follow a discriminant evolution (see figure 4).

C. Repetition index

From the audio inspection of the corpus of soccer broadcasts it has been noticed that during the highlight moments, or just before and after them, most of the sport commentators tends to repeat short words, such as ”goal”, the name/nickname of the player, or to lengthen in time vowels, for example the ”o” in the word ”goal”. Therefore, a repetition index has been studied as a feature to discriminate hot spots in soccer retransmissions. To the authors knowledge, the acoustic repetition index has not been used previously in similar applications.

To compute the repetition index the audio track is divided in 2 second length windows with a 50% of overlap. The central part of each window is selected as the narrow acoustic section of interest, and the normalized cross correlation between this section and the rest of the window is computed. Finally, the repetition index is stated as the mean of the N higher values of the cross-correlation (with $N = 5$).

Figure 5 shows the repetition index distribution for neutral events, highlighted events and goals. Although the distributions are mainly overlapped, the repetition index evolution in a single file (see figure 6) manifests that there is a correlation between the repetition index few seconds before and after the goals events. Therefore, the repetition index for each second (the resolution of the system) has been computed as the maximum value of that index in a 5 seconds length window centered in the second of interest.

VI. Soccer game summarization system

The soccer game summarization system is composed by two blocks: a highlighted event detector and a video generator. In the following sections both blocks are described.

A. Highlighted event detection algorithm

Once the acoustic features are computed the system elaborates a list with hot spots positions and their score, indicating the relevance of each of them.

The hot spot time position is determined based on the energy block feature. The positions are iteratively determined

by looking for the point with the maximum energy block value and masking the values around, to assure not to select the same hot spot more than one time. In this study we have considered a fix duration of each hot spot of 20 seconds, 10 seconds just before the position of the pick of the energy value and 10 seconds after.

Once the positions are determined, the score or relevance is computed based on the energy block feature and modified by the repetition index. We do not use entropy information since the entropy has been proved to be not discriminant.

The procedure to compute the score is as follows. For those instants which the repetition index is higher than a threshold $Th_{RepIndex}$, the score is directly the energy block value, since there is a strong correlation between the energy level and the relevance of the moment in the game. And for those instants which the repetition index is lower than a threshold $Th_{RepIndex}$, the score is assigned the 95% of the energy block value. The threshold is determined by the statistics of the repetition index, in particular:

$$Th_{RepIndex} = \mu_{RepIndex} + 1.5\sigma_{RepIndex} \quad (1)$$

where $\mu_{RepIndex}$ is the mean and $\sigma_{RepIndex}$ the standard deviation of the repetition index.

With this procedure, the points with high energy levels and high repetition index will remain as relevant events, but those points with not so high repetition index goes down in the relevance ranking. The reason for this modification is to re-order the relevance of the highlighted events according to the repetition index, to increase the relevance score of the goals, since it has been observed that the repetition index is high specially in the goal events.

B. Video soccer game summary generation

Soccer game summaries of different durations may be generated concatenating the video clips of the N most relevant hot spots, according to the score computed with the procedure explained in the previous section.

In this study we have not carried out any automatic processing to determine the duration of each individual hot spot, but we have empirically checked that a duration of 20 seconds is good enough to cover the most part on the relevant attacks in soccer games.

VII. Results

Experiments with the UEFA EURO subcorpus have been carried out in order to determine the performance of the proposed system. In particular, two figures of merit have been considered of interest for the applications to be covered by this system:

- The recall of goals (GR), defined as the ratio between the retrieved goals in the summary in front of the total number of goals in the match.

- The precision of the summary (SP), defined as the ratio between the relevant events of the summary in front of the total number of events of the summary.

Duration	Hot Spots	GR	SP
2 minutes	6	70.73	73.62
3 minutes	9	82.93	67.41
4 minutes	12	95.12	60.00

TABLE III
Recal of goals (GR) and precision of the summary (SP) for the UEFA EURO subcorpus.

The goal recall is the key figure in our design, because the perception of the quality of a summary by the most part of the final users of alert systems is highly influenced by the presence of the totality of the goals of that particular match. Professional sport editors working with a pre-selector of highlighted events are also more concerned by the certainty that all the goals would be pre-selected than by the presence of non highlighted events in the pre-selection.

Table III shows the results for different summary durations. The goal recall increases when more highlighted events are retrieved, achieving a 95% recall when selecting 12 events. This is a promising result for both target applications, specially because the duration of the summary may be reduced by selecting the video of the events with less than 20 seconds. Also, we have checked that the amount of events that must be selected to achieve a 100% recall in the UEFA EURO subcorpus is 25, a number very suitable for the application of a pre-selection of relevant events for sport editors.

The numerical precision figure is not outstanding, but it must be remarked that the labeled highlights were only those showed in table II. Spectacular dribbling or promising attacks not ending in a shot to goal are not labeled as highlights, therefore not countered in the precision figure. Informal test with human evaluators have corroborated than the precision perceived by the users is the adequated and higher than the numerical figure.

VIII. Conclusion and future work

The goal of this paper is to design a highlighted event detector for soccer games to cover two industrial applications of automatic summary generation: one application for final users of soccer alert systems and one application for professional sport editors. The main requirements of both applications are near real-time performance, language independence and a high recall of goals.

The designed system is based on very simple and fast computable acoustic features: the block energy to select the highlighted instants and a repetition index to refine the relevance score. The repetition index represents the correlation between a narrow acoustic section and the seconds just after and before it, in order to detect sections of audio where repetitions occur. Experiments on a corpus of soccer games retransmission with sport commentators have validate the design.

Future work will include to incorporate a new technique to determine each individual hot spot duration instead of

using a fix duration, for example based on scene change point detection.

ACKNOWLEDGMENTS

The authors would like to thank Jos Gregorio Escalada and David Cacenabes for their useful and interesting remarks about the proposed algorithm, and also for the time dedicated to manually check the goal and other highlight labels.

This study has been carried out in the frame of the Spanish government founded project "CENIT 2007-1012 i3media: Tecnologas para la creacin y gestin automatizada de contenidos audiovisuales inteligentes".

REFERENCES

[1] J. Assfalg, M. Bertini, A. D. Bimbo, W. Nunziati, and P. Pala. Soccer highlights detection and recognition using hmms. In *in Proc. ICME*, 2002.

[2] M. Baillie and J. M. Jose. Audio-based event detection for sports video. In *in Proc. CIVR*, 2003.

[3] F. Coldefy and P. Bouthemy. Unsupervised soccer video abstraction based on pitch, dominant color and camera motion analysis. In *in Proc. ACM International Conference on Multimedia*, 2004.

[4] H.-G. Kim, S. Roeber, A. Samour, and T. Sikora. Detection of goal events in soccer videos. In *Proceedings of the SPIE*, volume Volume 5682, pages 317–325, 2005.

[5] M. Kolekar and S. Sengupta. Semantic concept extraction for sports video for highlight generation. In *in Proc. MobiMedia conference*, 2006.

[6] J. Li, T. Wang, W. Hu, M. Sun, and Y. Zhang. Soccer highlight detection using two-dependence bayesian network. In *in Proc. ICME*, 2006.

[7] S. Marlow, D. A. Sadlier, N. O'Connor, and N. Murphy. Audio processing for automatic TV sports program highlight detection. In *in Proc. ISSC*, 2002.

[8] Y. Rui, A. Gupta, and A. Acero. Automatically extracting highlights for tv baseball programs. In *in Proc. ACM international conference on Multimedia*, pages 105–115, 2000.

[9] C. G. Snoek and M. Marcel Worring. Time interval maximum entropy based event indexing in soccer video. In *In Proc. ICME*, 2003.

[10] C. G. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25:5–35, 2005.

[11] D. Tjondronegoro, Y.-P. P. Chen, and B. Pham. The power of play-break for automatic detection and browsing of self-consumable sport video highlights. In *in Proc. SIGMM*, pages 267–274, 2004.

[12] B. T. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications and Applications*, 3, No 1, 2007.

[13] K. Wan and C. Xu. Efficient multimodal features for automatic soccer highlight generation. In *Proc. of the 17th Int. Conference on Pattern Recognition (ICPR)*, 2004.

[14] J. Wang, C. Xu, and E. C. ans Qi Tian. Sports highlight detection from keyword sequences using hmm. In *In proc. ICME*, volume 1, pages 599–601, 2004.

[15] Z. Xiong, R. Radhakrishnan, and T. AjayDivakaran, A. Huang. Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework. In *in Proc. ICASSP*, pages 632–635, 2003.

[16] S.-C. C. M.-L. S. M. C. C. Zhang. A decision tree-based multimodal data mining framework for soccer goal detection. In *in Proc. ICME*, pages 265–268, 2004.