Summarizing Soccer Videos without Detecting the Events

Ehsan Lotfi, Hamid Reza Pourreza

P.O.Box 91775-1111, Computer Engineering Department, IntelliSys Lab, Ferdowsi University of Mashhad,Iran

Summary
For automatic summarization of soccer videos one can do the detecting tasks at the beginning of the game and then proceed to record with events. The new method we have provided here does the summarizing without detecting the events, and provides us with a mechanism for control the time of output film. One of the tools we used to do this was distinguishing between the views of the goal and the field-center. For this we present a 3-phase algorithm. The experimental results show the high accuracy of the proposed algorithms and the very good summarized output films.

Key words:
Soccer video processing, Shot classification, Event detection, Soccer video summarization.

1. Introduction

As watching a football match needs a lot of time, many TV fans of sport competitions prefer to watch a summary of football games. To do automatic summarization many works have been produced in the past which are mentioned briefly in this paper. During a football match whenever the ball crosses the goal line and enters the goal, a goal has been scored. To recognize this goal we can detect the ball object in the screen and monitor its moving trajectory to be able to recognize the concept of a goal. Such methods use object-based features, and some papers have used this method, e.g. [1] that uses object-base features to recognize major events, and [4] that uses object trajectories and relations to do so. On the other hand, certain features may be used to recognize the major events; some of these features are slow motion, spectators' excitement, subtitles or other types of texts on the screen, etc. These features are extracted from sound and video sources and are called *cinematic features*. Some of the papers have only used these features for summarizing football matches, such as [5] that only uses sound to generate the summarized version, and [6] that uses the camera motion parameters to detect the major events of a match. In [9] the three parameters of (a) ratio of the number of pixels of the field grass to the total number of screen pixels, (b) the oblique line defined across the field, and (c) correlation of the size of an object to the size of the whole image. This method employs the Bayesian network to classify the shots and then uses the shots order to recognize goals, corner kicks, and attacks. Any of the articles have used a combination of these two methods, e.g. [7], [8], [10], [12]. Some cases the bit streams of MPEG files have been used, such as [13], [2]. According to the conducted studies, there are four general ways to recognize high level semantics such as scoring a goal in the videos of sport matches:

(i) Methods which use object-base features.
(ii) Methods which use cinematic features.
(iii) Methods which use the information contained in Mpeg bit streams.
(iv) Methods which use a combination of the above.

Cinematic features are divided into two categories: 1) visual features and 2) audio features. In general, if we determine all objects inside the image, including the ball, goals, players, etc. we will be able to recognize many events of the match by having an eye on FIFA laws and regulations but, this task needs a lot of time and money. In contrast, using cinematic features provides us with a good trade off between the volume of required calculations and preciseness of recognition of high level concepts.

As stated above all methods are after selecting the shots which contain the major events. So they do the summarizing operation based on recognition and selection of such events. The basis idea in this paper is to indicate that other than selection of good shots, provisions should be made for rejection of useless shots. As we all know most of the time the ball is in the middle of the field and the major events such as goals, corner kicks, penalties, etc. mostly occur in shots of field sides. Consider a summary that includes removal of all probably useless shots, such as shots of fans, close shots of coaches, shots of the midfield, etc. Such a summary will mostly cover the attacks and major events of the match (except for some fouls). The structure of this article generally explains the block diagram of Fig. 1, and in the end, the result of the proposed system performance evaluation is shown.

2. Overall Structure

The most important result obtained by accepting the

concept of Rejection is that we will not use Event Detection (ED) for summarizing. In other words, just like the other works in scientific journals which first detected the events of a sport match and then proceeded to prepare the summary; we will no longer do the ED but will reject the shots containing non-significant events and moments to summarize the video. The overall structure of the proposed plan is shown in Fig. 1.

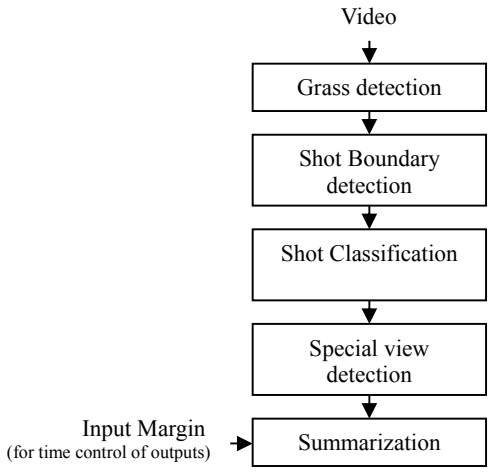

Fig. 1 The overall structure of the proposed plan.

2.1 Detecting the Play Area including the Grass Field

There two ways to detect the grass field. In the first method we should first determine in the HSI space those figures that show the color of various grass fields in various seasons and times. By the use of such figures we will be able to determine the grass field in an image; this is done off-line. But the second method includes extracting the area including the grass field in a startup style [12], which bears nice results. We have used the second method here.

2.2 Detecting the Shot Boundaries

A usual way to detect a shot on the video is to use the difference in the histograms of two frames. Only in videos of football due to having a single-color background in most frames this method is not useful in isolation; therefore, another criterion, i.e. difference of the percentage of pixels making the grass field in frame i, j, Gd(i,j), has been introduced in [12]. The results of implementing such a method for detecting the shot boundaries [12] are included in Table 1.

2.3 Classifying Types of Views

The types of views usually seen in a football video are classified into 4 groups:

1. Far views from midfield, showing an overall image of the midfield, Fig. 2.a.
2. Far views from field sides, Fig. 2.b.
3. Medium views from inside the field, where the camera has zoomed on full body of a person, Fig. 2.c.
4. View of the area outside of the field / closed view, showing the spectators or the upper body of a player, Fig. 2.d.

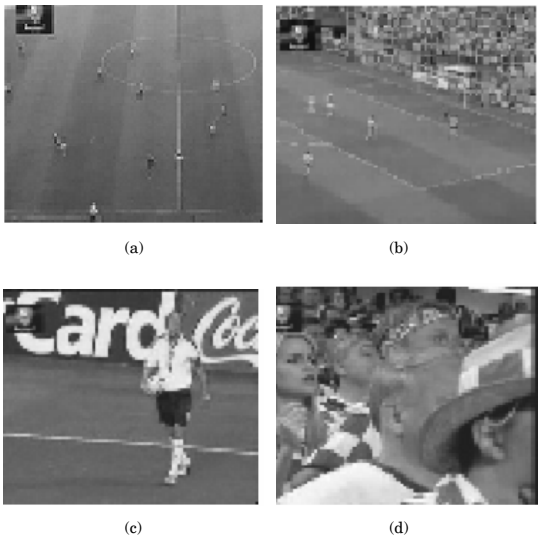

Fig.2 Types of views: a) Far-center, b) Far-side, c) Medium view, d) Out of field

Fig. 3 Segmenting the area including the grass field to distinguish between midfield and Far-Side views from the medium view

It is obvious that views 1, 2, and 3 are distinguished

from view 4 by the aid of G_i (grass colored pixel ratio in i_{th} frame). The distinguishing boundary was experimentally and primarily considered as 0.4. Here we present an algorithm for distinguishing between view 3 and views 1 & 2. If the area including the grass is segmented with ratios of 3:5:3 as shown in Fig. 3, and then calculate the percent of grass pixels for the 3 segments of E_1, E_2, and E_3 we will have G_1, G_2, and G_3. Afterwards, we can use the following linear formula and define a boundary to distinguish between view 3 and views 1 & 2.

$$G_a = \frac{3}{7}G_1 + \frac{4}{7}G_2$$
$$G_b = \frac{3}{7}G_3 + \frac{4}{7}G_2 \quad (1)$$

The said boundary will be known after one phase of learning. The pattern for distinguishing the views is shown in Fig. 4.

2.3.1 Distinguishing Long Views of Field-sides from Long Views of midfield

In the previous section the segment algorithm perfectly detects the long views. Now we propose a method for recognizing far-center or far-side views, which is based on an interesting feature of football videos, it being the fact that most long shots in football matches are taken by a main camera which is located at a fixed place in relation to the field. This feature results in regular and uniform views in the long-shot. For example, in most long shots either the longitudinal or the cross lines of the field are seen, the first in far-center and the second in far-side views. It is interesting that due to the fixed position of the central camera, the angle of the side line which is seen in the image has close or similar values in far-center or far-side views.

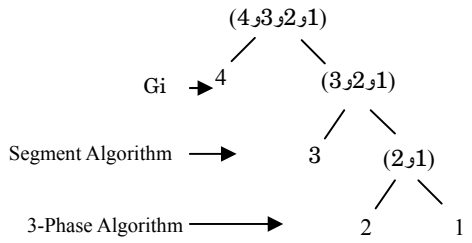

Fig. 4 Model for Distinguishing between the Views

To do so in a 3 phase method (Fig. 5) all objects inside the image are defined as closed lines and then the type of view is distinguished by extracting the formula of the line that encompasses the whole field object.

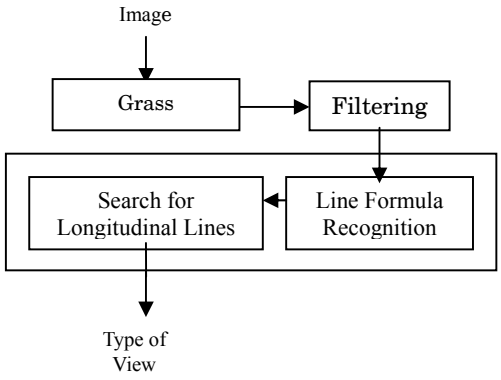

Fig. 5 The 3-phase method for recognizing the type of far view

Phase 1: The grass detection algorithm was described in the previous sections. In fact phase 1 was done in the previous sections and its output is now ready. The result of phase 1 is a binary image in a way that each pixel of grass field is shown with "1" and each non-grass pixel is shown with "ø". Phase 2: By using a high pass filter in the binary image all border lines are show. the high pass filter used here is a laplasin filter. Fig. 6 shows a sample of the output of phase 2. Phase 3: To detect the black lines in the white page we have made an algorithm which is based on Hough Algorithm. However, this algorithm has been optimized for our purpose here. Fig. 7 shows a sample of the results of grass border lines detection. Indeed detection of the two sorts of far views is done by defining an edge over the angle of the lines estimating the cross lines of the field. The said edge will be determined after one phase of learning. The final result is shown in Table 2.

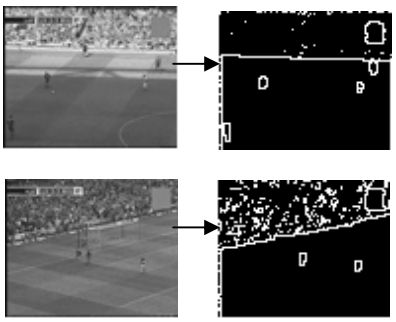

Fig. 6 Two sample of the result of applying Laplasin Filter to the output of grass detection phase

(a) (b)

Fig. 7 Detecting the Borderline of Grass field by the 3-phase algorithm (the group of lines estimating the grass borderline in *a* have been drawing with black in *b)*

2.4 Classifying the Shots

Shots which occur according to the change of close-up and in-field medium views are classified with the same title, i.e. Close-Up Shot and In-Field Medium Shots, and there is no difference between the meaning of change of view and the concept of shot about them. Yet for far views we have to distinguish between the meanings of a shot and change of view, because in long-shots the change of view from center field to a view of field sides will not necessary be a shot. Such a change may be done with a very slow motion of the central camera. So to be able to use such views (the goal and field-side views) we need to present a new and precise definition of long-shot which is useful in summarizing:

- Long-shot useful for summarizing: A shot in which enough views are from field-sides or the goals. In practice we should prepare statistical data from all views existing in a long shot, then we will have to give every long shot a Far-Side degree, in other words, if all of the long-shot is made of far-sides it should have a value of 1 and if all the long-shot is made of center field views then it should have the value of 0. The values between 0 and 1 should be given to shot including the same proportions of far-side and far-center views. Thus we allocate an amount to each long-shot indicating its far-sidedness.

3. Summarizing

After such processing efforts we will have all the relevant information about the beginning and ending times of a shot, type of the shot, and Far-Side degree of it which are all required for summarizing. To summarize we first need to provide the system with a boundary input. This input will define the useful and non-useful shots for summarizing purposes. For instance, if the value is 0.25 the long shots with Far-Side degree less that 0.25 (that is the percent of far-side views in them is less than 25%) are the shots which probably contain no major events user is looking for. These shots are rejected in the summarizing process.

So by defining this boundary the system will be able to distinguish between useful and non-useful long shots. The system then follows these instructions for summarizing:

- "If there are no useful shots between two consecutive long-shots both with zero far-side degrees, reject all shots between them including the two shots themselves."
- "If there is at lease one useful shot between two consecutive long-shots both with zero far-side degrees, reject all close-up and out-of-filed shots between them."(Optional).

3.1 Time Control of Output Films

As mentioned before, by defining a margin on the far-side rate of a long-shot, as the system input we can separate the useful and none-useful shots for the purpose of summarizing. The higher the margin the shorter the time of the summarized film (Fig. 8). The highest amount of the margin definable in theory is 1, the application of which to the system causes the summarized output to include only events for which at least one shot occurs which includes views of the goal such as corners, penalties, etc. The lowest amount of the margin is 0 (zero). By defining the input margin as 0 the input film will be seen in the output as well. The other margin amounts are between 0 and 1 that according to the opinion of various people results in a degree of weakness and strength in the output attacks.

4. Evaluation

To evaluate the system we have used more than 5.5 hours of football including 1 match of the World Cup 2006, 2 matches from the UEFA Champions League 2005, 1 match from the FA Premier League 2004, and more than 5 short clips from Euro 2004. The file format of the input films was non-compressed avi and the size of output films is 88×79.Table 2 shows the precision of the 3-phase method designed for detecting the type of long-shots from the field-side views. In the 3-phase method the only part that needs one phase of learning is that related to classifying and grouping the sidelines of the field. This has been achieved by using only 20 minutes of all films. In implementation of some of the sampled views of each long shot for allocating a far-sidedness rate to that shot, 5 was given to the system. This way the input margin of the system is 0.2 and inputs include 0.2, 0.4, 0.6, 0.8 and 1. The results of changing the input amount of the system and the duration of the output film are shown in Fig. 9. This is a proof of what we said in 3.1 section and reminds us of a pyramid structure. Table 3 is related to statistics of output films based on inputs of 0.2, 0.4, 0.6, 0.8 and 1.

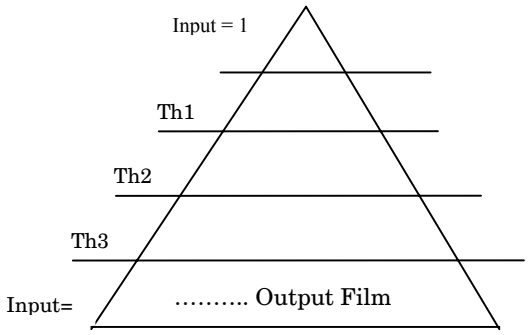

Fig. 8 Relation of Output Film Time with Input Margin (Th1>Th2>Th3).

Table 1: The result of Shot Boundary detection

# of Far-views	36	33	26
Correct	29	28	24
False	7	5	2
Accuracy(%)	81	85	92

Table 2: The result of Distinguishing Field-side from midfield views.

Length of clip(min:sec)	20:51	18:30	12:35
# of shot	144	133	71
Correct	122	188	65
False	19	11	4
Miss	3	4	2

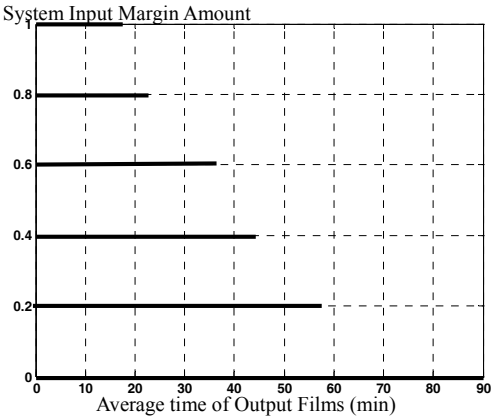

Fig. 9 Control of output film time is possible by defining the margin amount for presence of far-side views in each long-shot.

Table 3: statistics of output films based on inputs of 0.2, 0.4, 0.6, 0.8 and 1.0

	Goals	Penalties	Free kicks	Corner	Other attacks
Total	3	5	6	9	48
Rejected (Input=0.2)	0	0	0	0	4
Rejected (Input=0.4)	0	0	1	1	16
Rejected (Input=0.6)	0	0	2	1	16
Rejected (Input=0.8)	1	1	4	2	29
Rejected (Input=1.0)	1	2	4	4	33

5. Conclusion

In this paper we have presented a new method for summarizing football videos in which there is no need for detecting the events. This method is able to make a summary of attacks including goals, corners, penalties, shots, and a group of strong and weak attacks based on user request. The presented method is very fast and capable of more optimization and full real-time implementation.

References

[1] J. Assfalg, M. Bertini, A. Del Bimbo, W. Nunziati, and P. Pala, "Soccer highlights detection and recognition using HMMs," *Proc. IEEE Int'l. Conf. on Mult. and Expo (ICME),* Aug. 2002.

[2] K. A. Peker, R. Cabasson, and A. Divakaran, "Rapid generation of sports video highlights using the MPEG-7 motion activity descriptor," in *Proc. of the SPIE conf. on Storage and Retrieval for Media Databases,* vol. 4676, pp. 318-323, Jan. 2002.

[3] A. Gueziec, "Tracking pitches for broadcast television," *IEEE Computer,* vol. 35, no. 3, pp. 38-43, March 2002.

[4] V. Tovinkere and R. J. Qian, "Detecting semantic events in soccer games: towards a complete solution," in *Proc. IEEE Int'l. Conf. on Mult. and Expo (ICME),* Aug. 2001.

[5] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," in *Proc. ACM Multimedia,* 2000.

[6] R. Leonardi and P. Migliorati, "Semantic indexing of multimedia documents," *IEEE Multimedia,* vol. 9, no. 2, pp. 44-51, Apr.-June 2002.

[7] W. Zhou, A. Vellaikal, and C-C.J. Kuo, "Rule-based video classification system for basketball video indexing," in *ACM Mult. Conf.,* 2000.

[8] D. Zhong and S-F. Chang, "Structure analysis of sports video using domain models," in *Proc. IEEE Int'l. Conf. on Mult. and Expo (ICME),* Aug. 2001.

[9] Ming Lou,Yu-fei Ma, Hong-jaing Zhang,"Pyramidwise Structuring for Soccer Highlight Extraction," *Department of Computer Science University of Maryland*.

[10] B. Li and M. I. Sezan, "Event detection and summarization in American football broadcast video," in *Proc. of the SPIE conf. on Storage and Retrieval for Media Databases,* vol. 4676, pp. 202-213, Jan. 2002.
[11] H. Pan, P. van Beek, and M. I. Sezan, "Detection of slow-motion replay segments in sports video for highlights generation," in *Proc. IEEE Int'l. Conf. on Acoustics, Speech, and Signal Processing (ICASSP),* 2001.
[12] A. Eklin, A.murat Tekalp and Rajiv Mehrotra, "Automatic Soccer Video Analysis and Summarization," in *Proc.IEEE Int'l Conf ,2003.*
[13] R. Leonardi, P. Migliorati and M. Parandini, "Semantic Indexing of Sports Program Sequences by Audio-Visual Analysis," *in ICIP Conf., 2003.*

H.R. Pourreza finished his M.Sc. in Electrical Engineering (1993) and Ph.D. in Computer Engineering (2002) at Amirkabir University of Technology, Tehran. He is Assistant Professor at Computer Engineering department of Ferdowsi University of Mashad, IRAN. His research interests include image processing, machine vision and intelligent transportation systems.

Ehsan Lotfi received the B.Sc. degree in Computer Engineering, from Ferdowsi University of Mashhad, in 2006. His research interest includes video processing, fuzzy digital image processing, machine vision, hardware simulation and analysis. He is a researcher of Khorasan Science and Technology Park in Iran.