Structure in Soccer Videos: Detecting and Classifying Highlights for Automatic Summarization

Ederson Sgarbi[1] and Díbio Leandro Borges[2]

[1] Fundação Faculdades Luiz Meneghel,
Depto. de Informática, Bandeirantes - Pr, Brazil
sgarbi@ffalm.br
[2] BIOSOLO, Goiânia - Go, Brazil
dibio@terra.com.br

Abstract. We propose an automatic framework to detect and classify highlights directly from soccer videos. Sports videos are amongst the most important events for TV transmissions and journalism, however for the purpose of archiving, reuse for sports analysts and coaches, and of main interest to the audience, the considered highlights of the match should be annotated and saved separately. This procedure is done manually by many assistants watching the match from a video. In this paper we develop an automatic framework to perform such a summarization of a soccer video using object-based features. The highlights of a soccer match are defined as shots towards any of the two goal areas, i.e. plays that have already passed the midfield area. Novel algorithms are presented to perform shot classification as long distance shot and others, highlights detection based on object-based features segmentation, and highlights classification for complete summarization of the event. Experiments are reported for complete soccer matches transmitted by TV stations in Brazil, testing for different illumination (day and night), different stadium fields, teams and TV broadcasters.

1 Introduction

With the widespread availability of digital formats for video making, production and TV broadcasting, an important area of technological and scientific interest that has emerged recently is Video Processing. Video Processing is naturally attached to the broad research areas of Computer Vision, Image Processing and Pattern Recognition, although it poses specific challenges concerning domain knowledge and computational resources. Videos are produced as documentaries, news materials, advertisement, movies, shows, TV programs, and sports coverage. Indexing and retrieving such material efficiently require new techniques in content and semantic description, coding and searching unavailable nowadays.

Soccer videos are major products in that industry attracting millions of spectators worldwide. However after a live transmission some plays, or shots of the match carry more interest than others, for example the attacks which came closer

M. Lazo and A. Sanfeliu (Eds.): CIARP 2005, LNCS 3773, pp. 691–700, 2005.

to a goal and of course the goals if any of the two teams scored. Those would be the highlights of the match. Broadcasters have personnel just for producing a logging of a soccer match, which could be then used by analysts in TV programs or as a main source for annotation and further saving in archives. We propose an automatic framework to detect and classify highlights directly from soccer videos. The proposed solution reported here present new algorithms for soccer shot classification, object-based feature segmentation into attack playing fields (i.e. left and right), midfield and stadium audience, and highlights detection and classification. Experiments are shown for more than 7 hours of video, comprising 4 complete matches in different locations, time of playing and teams. More than 94.0 % of the highlights were correctly detected and classified (i.e. recall rate), and the final produced summary is presented with less than 9 min of video instead of the complete 90 min match (i.e. compression rate achieved with summarization is bigger than 90 %).

The following sections of this paper comment on related research found in the literature, present more details of the proposed approach, show and analyze a great deal of experiments in order to evaluate the performance of the system, and draw conclusions about the achievements at this point.

2 Related Works

Sports video summarization research has been a hot topic in the last five years. Reports found in the literature explores sports such as baseball, tennis, and soccer mainly but the list is growing [2]. Worldwide soccer is the main sport attraction, and since a complete match takes more than 90 min, summarizing it including only the highlights is a real necessity for broadcasters and video program makers.

In the literature different features are employed in the attempt to summarize soccer videos. Edges and color are used as features in [4] to detect and recognize the line marks of the field. Motion detection is also used to identify particular camera motion patterns, and along with line marks decide if the scene is part of a highlight or not. The tests shown by the authors use a small number of presegmented shots as input and check he correct identification by their system. Their solution rely very much on the line marks detection, and from our experience those features appear occluded and sometimes indistinguishable, especially in shots near the goal area (i.e. highlights) when the players are much closer to each other than in other shots.

A set of fuzzy descriptors is used in [1] to represent and classify the positions of players in the field. Hidden Markov Models are then trained with these descriptors to help identify some subset (penalty, free, and corner kicks) of highlights for classification. Their experiments show only 10 shots pre-segmented for each of the highlights considered and then tested. Tests to show the performance of the players position detection are not given. Identifying the highlights directly from the transmission is a crucial bottleneck in this application, which if not accomplished compromise the whole summarization task.

Other work to use HMMs is [8], however their proposed classification is into ”plays” and ”breaks” shots only. Motion vectors and color ratios are used as features and a combination of those are trained using HMMs to separate into the two classes: play and break. Their experiments include parts (not complete) of matches and accuracy achieved was around 85 percent. In order to summarize the event a classification into highlights and not only ”plays” is necessary.

A combination of color and texture features are used in [6], and [7] in order to identify the players' shirts and track them in a shot. Medium and short distance frames are shown with identification of players with those features. Their proposed system allows tracking of players in a fixed camera situation. From the point of view of a summarization task, classifying highlights, it seems one pre-processing tool yet, because the highlight can not be decided based on these features only.

A summarization soccer system based on cinematic and object-based features is presented in [3]. Color and a spatial ratio mask are used as object-based features in the identification of long, medium and short shots. A cinematic template checking for duration of a break, slow-motion replay and close-ups is a proposed procedure to detect and classify particular events in a match. The events, or the equivalent highlights they propose to detect are: a) goals; b) referee; c) penalty box. Their experiments shown have recall rates of 85.3 % for those mentioned events for 13 hours of soccer video. We argue in this paper that the detection of a referee as an event does not seem to be an interesting highlight of a match, and that their ([3]) penalty box detection relies very much on the line marks and in most attack plays this type of shot is cluttered with players. Actually even that a combination of those events could help identifying a highlight this will be particular to a broadcaster style (e.g. showing a referee in every highlight shot), and the final summary could miss many interesting highlights for not detecting the penalty box.

Our approach presented in this paper addresses those issues, and proposes a complete summarization system for soccer videos based on direct object-based features, and efficient and robust new algorithms for achieving it. We propose to identify highlights as action in the attack fields, not only goals, and we evaluate the performance of the system in more realistic and difficult conditions for 4 complete matches. By more realistic and difficult conditions we mean by using direct transmissions of matches from TV in different stadiums and light (time of the day) conditions. The rest of the paper describe the proposed approach, the experimental protocol, results and conclusions.

3 Structure of a Typical Soccer TV Transmission

Typical transmissions of soccer matches on TV are designed to give a constant view of the main action, close-ups of some shots, and usually when breaks occur some replays. Some broadcasters might even use many different cameras to fill with other viewpoints of the play. Figure.2 shows the three main categories of shots that are used: 1) Long distance shots, 2) Medium distance shots, and 3)

Short distance shots. Important pieces of the game are mostly shown as Long distance shots, since they give a better view of the whole action in a play because the dimensions of the field and number of players in the game.

Semantically we could point two types of plays happening according to developments towards a goal: 1) Action in the midfield, 2) Action in the attack fields (i.e. either right or left). Scoring goals are the main objective of the game, however since it is not so easy to score a goal in soccer highlights of a match will be when the action is placed in the attack fields, i.e. the regions outside the centre circle and closer to the goal areas.

An automatic system to detect and classify the highlights of a soccer video will have to first identify the Long distance shots, then parse the shots as actions in the attack fields (i.e. highlights) and in the midfield. Approaches considering tracking the ball are not robust in practice since it is a very small object in a long distance shot, it is partially occluded in most of the scenes because of the players and marks on the field. Model-based recognition of the goal area using the marks on the field suffer also from the clutter in the scene, and in some fields especially in rainy weather they become indistinguishable.

Our approach will be to propose object-based features to be able to segment the action happening in any of the attack fields, without having to track the ball or follow the marks on the field. As it is shown here with the experiments the solution proved to be very effective and robust to many of the situations encountered in soccer videos.

4 Framework of the Solution Proposed

A functional diagram of the soccer video summarization approach proposed is shown in Figure.1. The video stream is captured from the TV transmission and digitized in color frames, 30 frames per second. The three steps of the approach are: 1) Shot classification, 2) Object-based feature segmentation, 3) Highlights detection and classification. Details of each step are given in the next sections.

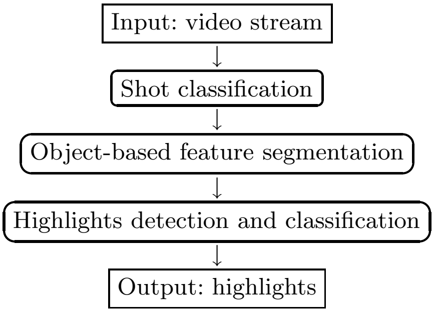

Input: video stream
Shot classification
Object-based feature segmentation
Highlights detection and classification
Output: highlights

Fig. 1. Main steps of the soccer video summarization approach proposed here

4.1 Shot Classification

Figure 2 shows the three main categories of shots in a soccer video. This step of the approach is aimed to classify the Long distance shots and pass them to the next step of the system. We devise the following algorithm for performing it:

i. Color frame is normalized in RGB, I = (R+G+B)/3;
ii. A histogram is computed for each frame and the dominant bin pixels are selected together with pixels belonging to the 10 % closer bins to the dominant one;
iii. Only frames which have at least 65 % of the pixels selected in the step ii. are picked;
iv. Sequences shorter than 100 frames are classified as Medium distance shots;
v. Sequences longer than 100 frames are classified as Long distance shots;
vi. Other frames which did not pass step iii. above are classified as Short distance shots.

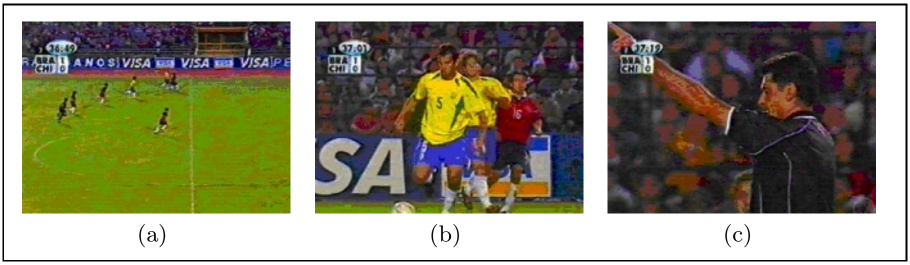

(a) (b) (c)

Fig. 2. Three categories of shots commonly found in a soccer TV transmission. (a) Long distance shot, (b) Medium distance shot, (c) Short distance shot.

4.2 Object-Based Feature Segmentation

This step has as input the long distance shots already classified earlier. The aim here will be to design and evaluate object-based features to be able to segment the frames into field and outside areas. Depending upon the concentration of these areas in a frame a decision procedure can be formulated to identify main action in the midfield, or attack fields to the left or to the right.

Upon analyzing the clutter in the long distance shots we devised the following procedure to perform segmentation into either field, or outside areas:

i. Find edges in the image (e.g. Marr-Hildreth filter);
ii. Place a grid of 16x16 cells upon the edge image;
iii. Try to fit a line in each cell by doing a Principal Components Analysis on their values;
iv. Cells will be marked either as "clutter", or "lines" depending on the residuals of the fits ([5]);

v. By checking neighboring cells for region consistency, clean (i.e erase) isolated cells marked as "clutter";

Figure.3 shows snapshots of this step of the approach. This step is performed on each frame of the sequence classified as long distance. Two consistent main regions are given as output of this stage, Figure 3.d) shows an example of the this output.

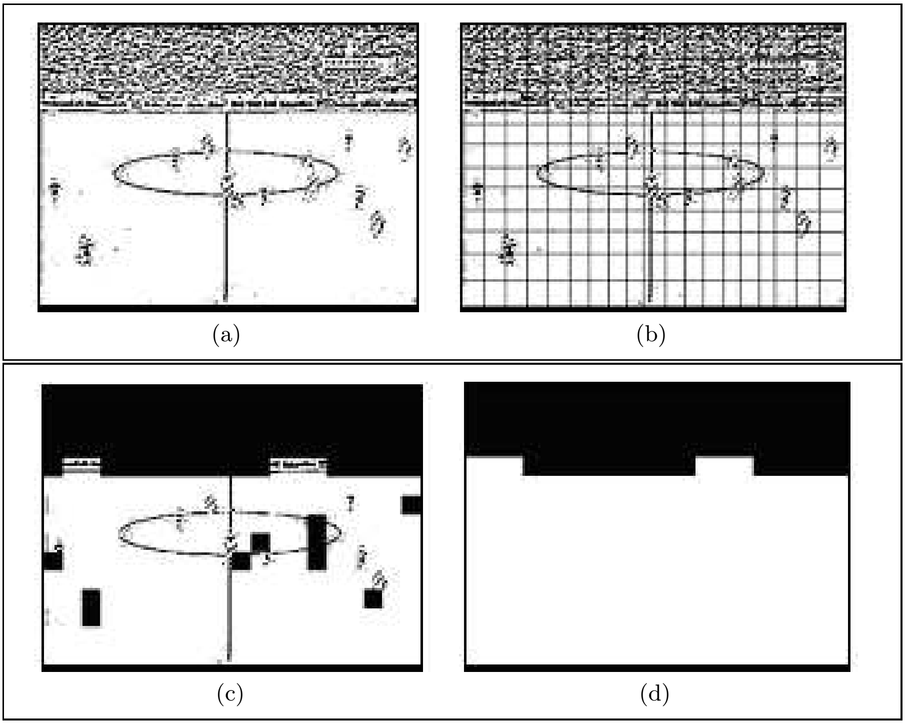


Fig. 3. Typical outputs of the object-based feature segmentation steps. (a) After the binarization, (b) With the grid superimposed to compute the eigenvalues, (c) After the decision on each window about a significant direction, (d) Result after cleaning inconsistent regions.

4.3 Highlights Detection and Classification

The final stage of the approach consists of the decision on highlights detection and classification based on the consistent regions given from the earlier step. The following algorithm performs this stage:

i. Place a 4x4 grid on the two consistent regions image given as input;
ii. Compute the density of black regions on the 16 cells of the grid;

iii. Higher density on the right side cells is classified as "highlight (attack on the right)";
iv. Higher density on the left side cells is classified as "highlight (attack on the left)";
v. Equal or higher density on the middle cells is classified as "not highlight";

Figure.4 shows snapshots of the final highlight classification, from image still with inconsistent regions (4.(a)), after cleaning inconsistent regions (4.(b)), and with grid for density computation placed upon it (4.(c)). The final classification of this example is "highlight (attack on the right)".

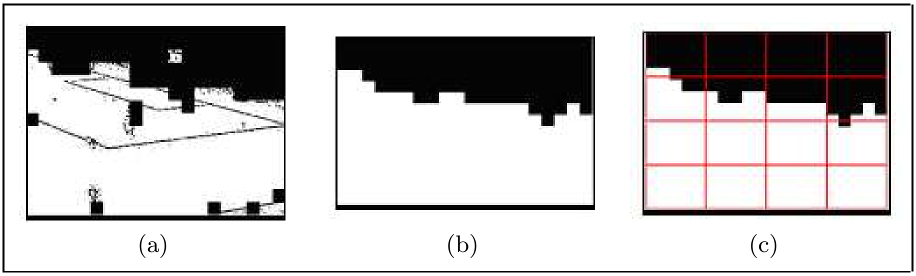


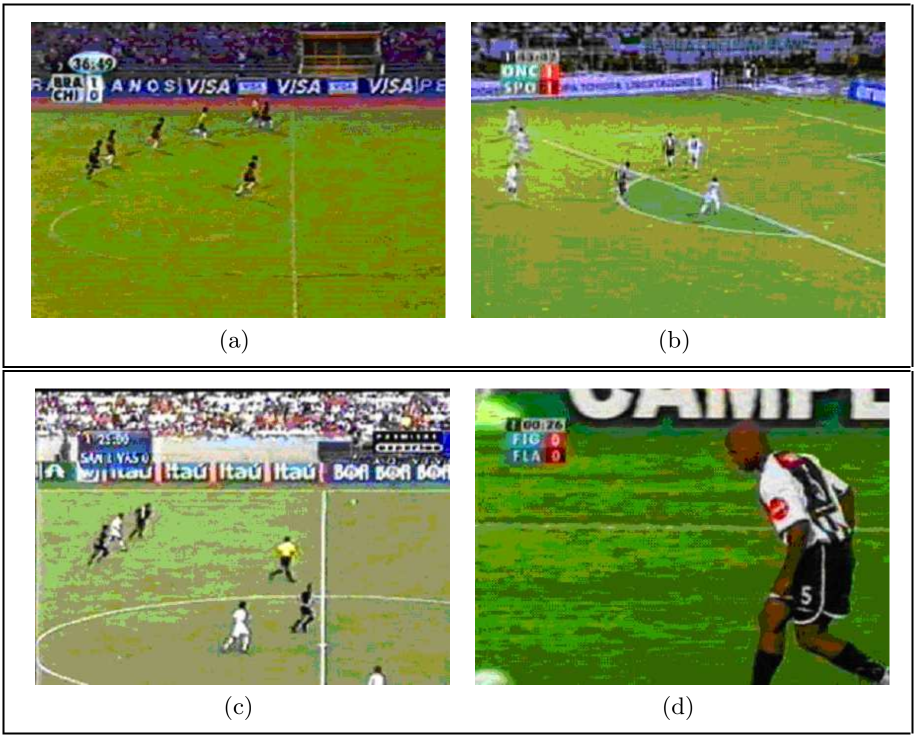
Fig. 4. Three steps on a shot detection and classification. (a) Region classification output before cleaning for inconsistent regions, (b) Final detection for classification, (c) Grid used for classification of highlight, showing an attack to the right.

Next section shows experiments performed for evaluating the approach presented here.

5 Experiments

Some of the works found in the literature of soccer video summarization rely either on the detection of the marks of field, or on cinematic features for processing motion. As we mentioned earlier in this paper those features are not robust in practice, since in most of the TV soccer transmissions (see Figure.5 for example shots) the marks on the field are difficult to recognize with efficiency necessary to perform such a task. On the other hand cinematic features are expensive to compute, and it brings too much burden in this task since 30 frames per second is the acquisition rate to process. The time when the match is played, the Stadium, i.e. the grass conditions of the field, the teams and the transmissions produced by different broadcasters pose realistic conditions to evaluate the approach. Figure.5 shows snapshots of soccer games to illustrate some of these conditions. In order to evaluate the approach proposed here we acquired four (4) complete soccer matches from TV transmissions in different situations: 1) Match G1 "Brazil x Chile", played at night, Conception Stadium, Chile (2004);

2) Match G2 "Figueirense x Flamengo", played at afternoon, O.Scarpelli Stadium, Brazil (2004); 3) Match G3 "Atletico(PR) x Botafogo(RJ)", played at afternoon, J.Americo Stadium, Brazil (2004); 4) Match G4 "Santos x Vasco", played at afternoon, B.Teixeira Stadium, Brazil (2004).



(a)

(b)

(c)

(d)

Fig. 5. Shots of four different soccer matches showing different situations to deal with. (a) and (b) are from matches at night, (c) and (d) during the day. All of them are set in different fields (stadiums).

Table.1 shows the compression rates achieved with the summarization approach proposed here for the different soccer matches mentioned G1, G2, G3, and G4. The number of highlights varies of course from match to match, however by detecting the highlights automatically only this information will be handled by TV program editors and placed for further annotation and indexing. The saving in time and computational resources is considerable since the average compression rate is 90.8 % for the 4 matches.

In order to have a ground truth for evaluating the performance of the system we annotated manually every shot in the 4 matches. Table.2 gives the results considering the expected input highlights from the ground truth. It is important to notice in this area that a false detection of a highlight is not of major concern

Table 1. Compression achieved by the soccer highlights detection algorithm in four (4) different games captured from TV transmissions. Data was acquired in full resolution, color, 30 frames per second.

	Input Frames	**Highlights**	**Compression**
G1	169200	8035	95.3 %
G2	166074	25958	84.4 %
G3	171513	16792	90.2 %
G4	168455	11611	93.1 %
TOTAL	**675242**	**62396**	**90.8** %

since they are to be passed to indexing and those could be cleared out. The recall rate, the relative success in detecting the expected highlights is of greater importance. On average the recall rate presented here is 94.6 %. Other works from the literature ([1, 3, 4]) report recall rates of less than 80 %. Although the test data are not the same we have used similar input size (4 matches), but in much harder conditions such as the field conditions and time of the play.

Table 2. Compression achieved by the soccer highlights detection algorithm in four (4) different games captured from TV transmissions. Data was acquired in full resolution, color, 30 frames per second.

	G1	G2	G3	G4	**TOTAL**
Input Highlights	190	463	280	280	**1213**
Correct	182	447	263	258	**1150**
Missed	8	19	17	22	**66**
False	82	100	157	169	**508**
Precision	68.9 %	81.7 %	62.6 %	61.4 %	**69.4** %
Recall	**95.8** %	**95.9** %	**94.0** %	**92.1** %	**94.6** %

By analyzing the results we could notice that many of the False highlights detected were due to some texts and logos appearing on the screen during the transmissions. They are usually placed by the broadcasters on either side of the screen cluttering the scene nearby the considered attack fields by the algorithm. Regarding the highlights missed by the system they were mainly due to sudden change of cameras during transmission, cutting from a long distance shot to either medium or short ones. Not many broadcasters use this, and the missed ones were below 5 %, however it would be a point to explore further with more experiments.

6 Conclusions and Future Works

In this paper we have presented an automatic system to detect and classify highlights of soccer videos. A complete summary of the match is achieved making it possible for practical use for annotation, indexing, and video retrieval. We have set a test protocol which includes more than 7 hours of soccer video, i.e. 4 complete matches, with a great variety of circumstances to evaluate the system performance. The summaries obtained produced a compression of 90.8 % from the input data, and a recall rate for the highlights of 94.6 %. This is a higher rate than others seen in the literature [3]. The successful results in such conditions allow us to explore realistic further possibilities of a fully automated soccer video summarization. The missed highlights and the false detected ones from our experiments were mapped to other features to be explored in future work, they are the text and logos appearing during transmission that should be dealt with, and sudden change of cameras in some shots. We are exploring these research lines in our group.

Video processing is an area of growing demand in research and development nowadays. It is a truly research area of Computer Vision and Pattern Recognition with well defined domains shaped with data available and industry demands. The work presented here could be extended to other sports videos as well, since as it was shown exploration of the knowledge of the game is important to predict where the main action will be happening.

References

1. J. Assfalg, M. Bertini, A. Del Bimbo, W. Nunziati, and P. Pala. Soccer highlights detection and recognition using HMMs. In *Proc. IEEE Int. Conf. Mult. and Expo. (ICME)*, pages 825-828, 2002.
2. S. Chang. The holy grail of content-based media analysis. *IEEE Multimedia*, vol.9, pages 957-962, June 2002.
3. A. Eking, A.Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. *IEEE Transactions on Image Processing.* vol.12, n.7, pages 796-807, July 2003.
4. Y. Gong, L.T. Sin, C.H. Chuan, H.J. Zhang, and M. Sakauchi. Automatic parsing of TV soccer programs. In *Proc. IEEE Int. Conf. Mult. Comput. Systems*, pages 167-174, 1995.
5. F. Szenberg. *Acompanhamento de cenas com calibração automática de câmeras.* Doctorate Thesis (in Portuguese), Departamento de Informática, PUC-Rio, Rio de Janeiro, Brasil, 2001.
6. N. Vandenbroucke, L. Macaire, C. Vieren, and J. Postaire. Contribution of a color classification to soccer players tracking with snakes. In *Proc. IEEE Int. Conf. Systems, Man, and Cybernetics*, pages 3660-3665, 1997.
7. N. Vandenbroucke, L. Macaire, and J. Postaire. Color image segmentation by pixel classification in an adapted hybrid color space. An application to soccer image analysis. *Computer Vision and Image Understanding.* vol. 90, pages 190-216, 2003.
8. L. Xie, S.F. Chang, A. Divakaran, and H. Sun. Structure analysis of soccer video with Hidden Markov Models. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP).* pages 4096-4099, 2002.