ALGORITHMS AND SYSTEM FOR SEGMENTATION AND STRUCTURE ANALYSIS IN SOCCER VIDEO

Peng Xu, Lexing Xie, Shih-Fu Chang
Dept. of Electrical Engineering,
Columbia University
New York, New York 10027, USA
{pengxu, xlx, sfchang}@ee.columbia.edu

Ajay Divakaran, Anthony Vetro, Huifang Sun
Mitsubishi Electric Research Lab
Murray Hill, New Jersey 07974, USA
{ajayd, avetro, hsun}@merl.com

ABSTRACT

In this paper, we present a novel system and effective algorithms for soccer video segmentation. The output, about whether the ball is in play, reveals high-level structure of the content. The first step is to classify each sample frame into 3 kinds of view using a unique domain-specific feature, grass-area-ratio. Here the grass value and classification rules are learned and automatically adjusted to each new clip. Then heuristic rules are used in processing the view label sequence, and obtain play/break status of the game. The results provide good basis for detailed content analysis in next step. We also show that low-level features and mid-level view classes can be combined to extract more information about the game, via the example of detecting grass orientation in the field. The results are evaluated under different metrics intended for different applications; the best result in segmentation is 86.5%.

1. INTRODUCTION

Mining information in video data becomes an increasingly important problem as digital video becomes more and more pervasive. And video structure analysis is an important sub-problem as the basis of further detailed processing. In this paper, we address the problem by segmenting soccer video into basic semantic units: *play* and *break*. Our results from play-break segmentation can be used as a foundation for event classification, summarization, and browsing. Applications of these techniques are very useful for both professional and general users.

Most prior works on sports video analysis and video segmentation are using shot as the element for analysis. Gong *et al* [2]classified key-frames of each shot according to their physical location in the field or the presence/absence of the ball. Several other works[4][7] analyzed tennis or baseball video using canonical scene (such as tennis serve or baseball pitch) detection, motion analysis and object-level filtering. In video segmentation, clustering of key-frames and integration with another modal[5] has been used to reveal scene-level structure. However, such approach is often ineffective for sports video due to frequent errors in shot detection, and the negligence or mismatch of domain-specific temporal structure. In soccer video, for example, each play typically contains multiple shots with similar color characteristics. Simple clustering of shots would not reveal high-level play transition relations. Soccer video does not have canonical views (e.g., pitch) indicating the event boundaries.

So instead of using the shot-based framework, we adopt a new framework, where frame-based domain-specific features are classified into mid-level labels through unsupervised and supervised learning, and temporal segmentation of the label sequences is used to automatically detect high-level structure. Moreover, fusion among multiple label sequences based on different features can be used to achieve higher performance.

Diagram of our system is shown in figure 1. A domain-specific feature (grass-color-ratio) is used to classify frames into 3 kinds of views according to video shooting scale. Then the sequence of view labels is processed to reveal play/break structure of the game. In the learning phase, this unique feature is manually identified and the grass color detector is learned through unsupervised learning from the initial segment of each video. View labeling of frames and segmentation of label sequences are trained through supervised learning and their core algorithms are invariant to new videos. Note in our system, replays will be classified as plays if play activities are shown in replays. Commercials will be classified as breaks since they do not have play activities. Individual components and the overall results are evaluated under different metrics intended for different applications; the best results: 92% for view classification, 88.5% for angle detection, and 86.5% for play-break segmentation.

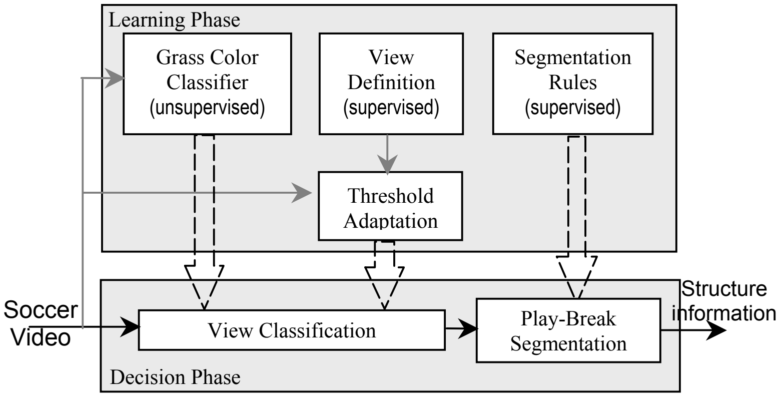

Figure 1. Block diagram of play-break segmentation in soccer video

In section 2 we discuss specific structure of soccer video. In section 3 we describe proposed algorithms and results for grass detection, view classification and grass-orientation analysis. Section 4 includes the algorithms and results for play-break segmentation. Conclusions are given in section 5.

2. STRUCTURE OF SOCCER VIDEO

In this section, we present some computable observations on soccer video, namely, the three types of views, and the correlation of view transitions with the play-break structure.

2.1 Shooting Scale and Views

There are three main types of views under common camera setup in soccer video production, named as follows: *global, zoom-in,* and *close-up*. Their counterparts in the general video making[1] terms are long shot, medium shot, and close-up, respectively. Semantically, these view types differ in their shooting scale, and this difference is usually reflected as the ratio of green grass area in soccer video, as shown in figure 2.

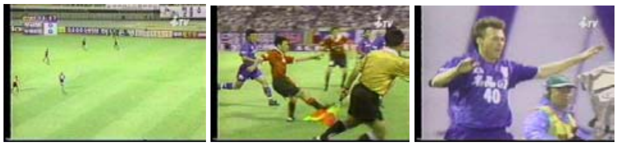

Figure 2. Three kinds of views in soccer video (left to right: global, zoom-in, close-up) global view has the largest grass area, zoom-in has less, and close-ups has hardly any

As we can see, a large amount of grass pixels are present in global view, there is some grass area in zoom-in view, while there is hardly any grass area in the close-up view (including cutaways and other shots irrelevant to the game without grass pixels). Hence, we use the grass area ratio feature for view classification in Section 3.

2.2 Temporal Syntactic Structure

Soccer game has two kinds of semantic states: play and break[3]. The ball is in play when it is within the boundaries of the field and the game is not stopped by the referee; break is the compliment, i.e. when the ball is wholly out of the boundaries, or the referee stops the game.

Compared to other kinds of sports video such as baseball or tennis, soccer game does not have a canonical scene for certain semantic events, as opposed to the serve scene in tennis[4] and pitch scene in baseball[7]; moreover, soccer is also characterized by a relatively loose time structure in the sense that play/break transitions or highlights of the game (such as goal, corner kick, shot, etc) is happening in a sporadic manner. However, there are some production rules that sports video-makers usually follow[1]. Producers typically aim to

- Convey the global status of the game
- Closely follow actions in the field

In order to meet these objectives, we have observed in soccer games that:

- During the play, it mostly stays in global view to convey the whole status of the game; interrupted by short zoom-ins or close-ups to follow up the players' action.
- During the break, zoom-ins and close-ups tends to be the majority as they can effectively show the cause and effect of the break (such as why a foul would happen, its consequence, etc.).
- A transition between play and break is usually within some time range, if not perfectly aligned, with certain transition of view.

These models about timing and transition are used in section 4 to process view labels and generate play/break segmentation.

3. FEATURE EXTRACTION AND VIEW CLASSIFICATION

As described above, the grass area ratio (i.e. the number of detected grass pixels to the frame size) is computed from the DC-image of each I-frame. The DC-images are extracted by parsing the MPEG-1 video stream without full decoding; then grass pixels are detected according to the grass detector; and view labels are obtained by quantizing the grass-ratios into 3 levels according to appropriate threshold values adaptively set in the initialization phase.

3.1 Learning Color-Based Grass Detector

Usually global and zoom-in view consist of a major part of the whole video, so it is reasonable to expect grass color to be the dominant color. To handle possible variations of lighting and filed conditions, we adaptively determine the grass color of each video by using a set of randomly drawn frames from the initial segment of the video. In our experiment setup, 50 frames are randomly drawn from a 5-minute clip (600 DC-I-frame pool), and the histogram of the hue component is added up over these 50 frames. We pick up the peak of this cumulative hue histogram over 50 frames, and use the corresponding hue value as the grass color.

This experiment is repeated 8 times over the same initial video segment to compute the mean and variation range of the hue value, which are used in the subsequent grass detector for this video. We also computed grass values from a segment that is 30 or 20 minutes apart from the original segment, and found that the variation of grass-hue values over time is smaller than the variation within the initial segment. So we assume any change of the grass-hue value due to lighting and field conditions over time is gradual, and thus can be periodically updated by automatic unsupervised learning during a long video program.

Although grass-hue values vary from 0.15 to 0.25 on a [0,1] scale across different games, the unsupervised learning method using the initial segment of each new video can be done when processing a new video program. By doing so, the grass detection accuracy has been quite consistent and satisfactory over different games.

3.2 View Classification

As discussed in Section 2.1, grass-area-ratio correlates very well to view type. We compute the ratio of number of grass pixels to the frame size (DC image from MPEG), and classify each frame to 3 types, corresponding to the 3 kinds of views. The classification is based on simple thresholding. We use heuristic values (0.1 and 0.5) as the initialization of the two thresholds, and then adjust the thresholds to the minimum values in the histogram of the grass-area-ratio feature of a 5-minute-long training video segment. The minimum point is searched within ± 0.1 of the initialized values. Such adjustment can be done for each new video program or for a class of video programs which are assumed to have similar production styles.

3.3 Grass Orientation Classification

Other useful features can be extracted from the detected grass area and view type. Grass orientation, together with the information about the camera setup, can be used to infer the approximate location of the field. Assuming the camera is following activities in the field, the location of the activity in the field can be inferred. We can also infer other information such as speed and type of movement in the game. An example showing the grass-stripe orientation can be seen in the left image in Fig. 2.

To ensure there is sufficient grass area in the image, we compute grass orientation only from DC images that have been classified as *global view*. Within each global view frame, we first compute gradient vectors $(g_{x,}g_y)$ of the grass using a Sobel edge mask on the luminance channel. Next, 50% of the grass pixels with the largest gradient value $G = |g_x| + |g_y|$ are declared as edge in the DC image. Grass orientation is computed as the average direction of the gradient vectors of all edge pixels in the grass area. Finally, average angle $\overline{\alpha}$ is compared to two thresholds θ, and $-\theta$ to classify each frame to right, middle, or left. Here the threshold θ indicates the range of the *Middle* class, as the criteria for "middle" will actually vary across human viewers. We set $\theta = 5^\circ$ in this experiment.

3.4 Experimental Results

The accuracy of *view classification* over 50 frames randomly drawn from 4 different clips is presented in Table 1. These clips belong to 3 different games (KoreaA and KoreaB are from the same game, 30 minutes apart) from different channels and countries. They represent sufficient diversity and variation. Note the performance is quite satisfactory considering the variation of the production styles and the simplicity of the feature and classification rules. Also note the majority of the errors are due to the breakdown of the assumption of the correlation between the grass area feature and the view, rather than the classification algorithms we developed. For example, zoom-in is sometimes shot with a grass-background, and the area of the grass background sometimes is even larger than that of some global shots.

Clip		KoreaA	KoreaB	Espana	Argentina
Thresholds	Th1	0.1744	0.1777	0.1348	0.1534
	Th2	0.5698	0.5272	0.4289	0.5074
Accuracy		92%	84%	82%	80%
Errors Due to Model-breakdown		6%	12%	16%	16%

Table 1. Accuracy of view classification: Th1 and Th2 are thresholds between close-up and zoom-in, zoom-in and global, respectively.

The set of frames used in evaluating *grass orientation* are randomly drawn from the clips KoreaA and KoreaB, where grass stripes are present. And then orientation classification is done on the frames that both the human viewer and the classifier agree as *global*. The accuracy of orientation classifier on these two clips are: 88.5% and 85.0%, respectively. Note this performance is quite good considering the noisy gradient in thumbnail image.

Almost all errors in these two classification algorithms come from the confusion of the middle class (zoom-in view or mid-field) with other two classes. This is a reasonable consequence in such 3-class classification tasks in 1-D feature space. And strictly speaking the labels are subjective to some extent, which introduces another level of confusion. For example, it is often hard to tell "mid-field" from "slightly left"; and zoom-in from close-up (especially if the foreground object is large).

Fuzziness in the classification stage and errors due to model breakdown will be taken into account in segmenting the label sequence, which will be described next.

4. PLAY-BREAK SEGMETATION

With the view labels obtained above, we proceed to segmentation for detecting play/break boundaries from the view label sequence. We also present performance under different evaluation metrics intended for different applications.

4.1 Segmentation of View Label Sequences

A sequence of view labels obtained from section 3 first goes through morphological post-processing to remove outliers and join adjacent segments, then a set of rules described below is applied to determine whether each view segment corresponds to a play.

These rules come from timing and transition observation in Section2.2. Here we assume play/break boundaries align with view transition. The main idea is to classify long, consecutive global view sequences as plays, consecutive close-up view sequences as breaks. At the same time, we try to resolve the fuzziness in short segments or zoom-in view by considering the majority labels in the neighborhood of the current segment.

Segmentation Rules:

1. Long segments of global view and close-up are classified as play and break, respectively. This decision is made according to the timing assumption in soccer games, and the threshold to determine "long" segments is chosen as $mean(T_i) - 1.3 * std(T_i)$, where T_i is the duration of all view segments from view classification output. This threshold actually correspond to 10% cut-off point if T_i observe a normal distribution.
2. Then the remaining segments goes through the following voting rule to exploit timing and continuity constraint.

Neighborhood voting rule: we count each kind of labels in the left and right neighborhood with equal duration of the current segment. And the outcome would fall into one of the following three situations:

Majority means more than half of the neighbors bear the same label; *Dispute* denotes the situation that all the left neighbors have identical label, and so do the right neighbors, but labels on the left and right are not the same; *Disagreement* means neither of the two situations above is true, i.e. no label represent more than half of the neighbors and at least on side of the neighbors are not in consensus.

3. Global view is classified as *play*, while the other two kinds of view are classified as *break* unless one of the

following conditions holds true in "neighborhood voting":

- Current label is *close-up* and the neighborhood *Majority* are *global*, then this *close-up* segment is merge as play.
- Current label is *global* and the neighborhood *Majority* are *close-up*, then this *global* segment is merge as break.
- If the current label is *zoom-in* and the neighborhood has a Majority label, then this segment follows the Majority (i.e. classified as play if majority is global, and vice versa)
- If the current label is *zoom-in* and the neighborhood is in *Dispute*, then the current segment follows the left neighbor. This accounts for the common production style that zoom-in view of the acting player is usually shown before he/she is going to start the next play; and the play often ends with close interaction between players that is often catched by a zoom-in view.

4.2 Evaluation Schemes and Performance

The segmentation results are evaluated under the following criteria according to different types of prospective applications:

1. *Global accuracy*- This is the ratio of the correctly-classified duration (play vs. break) to the total time of the test clip.
2. *Coverage of play*- We report *miss* when a certain play in ground-truth is totally misclassified as break, and we report *false alarm* if a certain play in segmentation result should be break in ground-truth.
3. *The start time of play or break*- This is to evaluate whether a counterpart in segmentation result can be found for each starting point of play or break in ground-truth within an ambiguity window of 3 seconds.

Criteria 1 and 2 are suitable for cases when the statistical status of all segments in the video is important. For example, some systems need to know the number of or the time percentage of plays/breaks. The 3rd criterion is useful when detection and timing information of each true play or break is important. For example, users may request to see a specific play from the beginning to the end. Tolerance for errors in missing the play or errors in the boundary timing is lower in this case.

The four 5-minute clips used here are the same as those used in section 3. Segmentation rules are applied to the label sequences coming from the view classifier as described in section 3.2.

Clip Name		KoreaA	KoreaB	Espana	Argentina
Global Accuracy		80.4%	86.5%	71.2%	67.3%
Play Coverage	Total	8	6	5	9
	Miss	1	1	0	0
	False Alarm	1	1	0	0
Play Start	Total	7	5	5	9
	Miss	3	1	3	5
Break Start	Total	7	5	5	9
	Miss	3	1	1	5

Table 3. Results of play/break segmentation

Errors are mainly caused by the breakdown of the assumption that a play-break transition corresponds to a transition of view. The two completely missed plays (the 2nd row in "play coverage") are both very short (less than 5 seconds) and are not accompanied by a transition of view. Some of the misses in play/break starting point is because the time of view transitions are either ahead or delayed so that they fall out of the 3-second ambiguity window. Errors are more prominent in the game named Argentina because the producer in this game tends to have lots of zoom-ins and close-ups during a play so that it is very hard to distinguish them from those actually correspond to short breaks. From the above results, we can notice that most plays and breaks are correctly detected (see "play coverage"), while the accuracy of the boundary timing can be further improved for some games. However, such timing errors may not be critical in applications in which accurate start/end times of the segments are not crucial.

5. CONCLUSIONS

We have presented new algorithms for analyzing the high-level structures and extracting useful features from soccer video. Specifically, we adopted a frame-based label-sequence processing framework for play-break segmentation. By exploring the unique domain structures in soccer video, we used a simple, but effective grass area feature to map sampled frames to mid-level labels (global, zoom-in, and close-up). We developed effective rules for segmenting plays/breaks from the label sequences. When tested over diverse programs from different sources, our system achieved very good results. We are currently seeking ways to eliminate errors caused by model breakdown, to extend current work by including complementary features such as motion or audio, and incorporating more advanced statistical learning tools.

6. REFERENCES

[1] Bob Burke, Frederick Shook, "Sports photography and reporting", Chapter 12, in *Television field production and reporting*, 2nd Ed, Longman Publisher USA, 1996

[2] Y. Gong et al "Automatic parsing of TV soccer programs", In *Proc. IEEE Multimedia Computing and Systems*, May, 1995, Washington D.C

[3] Soccer Terminology http://www.decatursports.com/soccerterms.htm

[4] G. Sudhir; J.C.M. Lee,; A.K. Jain, "Automatic classification of tennis video for high-level content-based retrieval", in *Proc IEEE International Workshop on. Content-Based Access of Image and Video Database*, Jan, 1998, Bombay, India

[5] H. Sundaram and S.F. Chang. "Determining computable scenes in films and their structures using audio visual memory models", *ACM Multimedia* 2000, Oct 30 - Nov 3, Los Angeles, CA

[6] P. Xu, et al "Algorithms and systems for high-level structure analysis and event detection in soccer video", *ADVENT Tech Report No. 111 , Columbia University*, June 2001. (http://www.ctr.columbia.edu/advent)

[7] D. Zhong and S.F. Chang, "Structure Analysis of Sports Video Using Domain Models", *IEEE Conference on Multimedia and Expo*, Aug. 2001, Tokyo, Japan