Exciting Event Detection Using Multi-level Multimodal Descriptors and Data Classification

Shu-Ching Chen, Min Chen
School of Computing & Information Sciences
Florida International University
Miami, FL, USA
{chens,mchen005}@cs.fiu.edu

Chengcui Zhang
Department of Computer & Information Science
University of Alabama at Birmingham
Birmingham, AL, USA
zhang@cis.uab.edu

Mei-Ling Shyu
Department of Electrical & Computer Engineering
University of Miami
Coral Gables, FL, USA
shyu@miami.edu

Abstract

Event detection is of great importance in high-level semantic indexing and selective browsing of video clips. However, the use of low-level visual-audio feature descriptors alone generally fails to yield satisfactory results in event identification due to the semantic gap issue. In this paper, we propose an advanced approach for exciting event detection in soccer video with the aid of multi-level descriptors and classification algorithm. Specifically, a set of algorithms are developed for efficient extraction of meaningful mid-level descriptors to bridge the semantic gap and to facilitate the comprehensive video content analysis. The data classification algorithm is then performed upon the combination of multimodal mid-level descriptors and low-level feature descriptors for event detection. The effectiveness and efficiency of the proposed framework are demonstrated over a large collection of soccer video data with different styles produced by different broadcasters.

1. Introduction

With the advent of large digital storage capabilities for broadcast sports videos, the challenges of efficiently indexing and searching in these video data sets become more acute. Sports videos are usually better structured compared to other types of videos such as home videos because they are under game-specific rules and regulations [19]. In addition, field-sports such as soccer, basketball, baseball, hockey, and rugby constitute the majority of sport videos. One common characteristic among all the field-sports is the restricted playfields with a defined layout. Several fixed cameras are installed above the playfield to capture sufficient game details.

Video events/highlights are often defined as the generically interesting/exciting events that may capture user attentions [11]. Most studies in video detection are game-specific as video events are innate game-specific. Domain knowledge is heavily used to infer the detection rules for such events [18]. There are also works on general video highlight extraction by replay extraction, motion extraction, and general analysis of audio and structures [9][10]. However, only coarse highlights can be extracted by these algorithms, while the detection of more game-specific events is more demanded by general users. Since a general algorithm that can detect the events in sports videos is hard to attain, we narrow our focus of this paper to a special kind of sports videos – the field-sports videos such as soccer, and propose an event detection framework with both generic and semi-generic multi-level multimodal descriptors.

The proposed video event detection framework is shot-based, follows the three-level architecture [5], and proceeds with three steps: low-level descriptor extraction, mid-level descriptor extraction, and high-level analysis. Low-level descriptors, such as generic visual and audio descriptors are directly extracted from the raw video data, which are then used to construct a set of mid-level descriptors including the playfield descriptor (field/grass ratio in soccer games), camera view descriptors (global views, medium views, close-up views, and outfield views), corner view descriptors (wide corner views and narrow corner views), and excitement descriptors. Both of the two modalities (visual and audio) are used to extract multimodal descriptors at low- and mid-level as each modality provides some cues that correlate with the occurrence of video events.

Low-level descriptors, though they can be extracted automatically, are generally not sufficient to support video event detection due to the "semantic gap"

between low-level features and high level subjective concepts. Heuristic rules can be used to partially bridge this semantic gap. For example, in [16], audio keywords are used to detect soccer events including foul, free kick, goal, etc., by applying a heuristic-rule based mapping. However, a set of thresholds need to be tuned in the heuristic rules, and some of the rules are biased because they are constructed based on the limited observations. For example, the rule "double-whistling indicating a foul" does not work for many other soccer videos. Mid-level features are constructed upon the base of low-level features and are expected to overcome the amount of variations in low-level features. As mentioned earlier, in this paper, a set of mid-level features describing camera view types, field ratio, level of excitement, and corner view types are proposed. Among these four features, the first three mid-level descriptors are generic descriptors for field-sports videos as they are not game specific. The latter descriptor is only semi-generic because while it is very useful in identifying corner events in soccer videos, it is a less important indicator of events in other types of field-sports like basketball and baseball.

The object features, such as object type and motion trajectories [12], if acquired, can offer even more information by allowing the direct reasoning of certain soccer events like goal attempts, kick-off, etc. However, the extraction of the object-level features is usually time-consuming and computationally expensive. Therefore, we propose to use a selective mixture of low-level descriptors and mid-level descriptors in the high-level analysis layer. For the inference of high-level events, hard-wired procedures are efficient in capturing the relations between the features and the events, but they are not generic [19]. Learning based statistical models, such as the Dynamic Bayesian network [14] and the hierarchical HMM [15], are chosen because it is difficult to intuitively define the relations between features and events.

In this paper, we present a new framework for exciting event detection, where each video shot is labeled by the set of low-level and mid-level descriptors and the events are inferred using learning based classification models such as the decision tree logic. Unlike the popular three-layer model [5] in which low-level features are only used to construct mid-level descriptor but discarded in the high-level analysis layer, our proposed framework takes both levels of multimodal descriptors in the event classification phase. Two typical exciting events, "corner events" and "goal events," are selected for experiments.

The remainder of the paper is organized as follows. Section 2 presents the procedure of extracting low- and mid-level descriptors. The discussions of high-level analysis are presented in Section 3. Experiments and discussions are given in Section 4. Section 5 concludes the paper.

2. Multi-level multimodal descriptors extraction

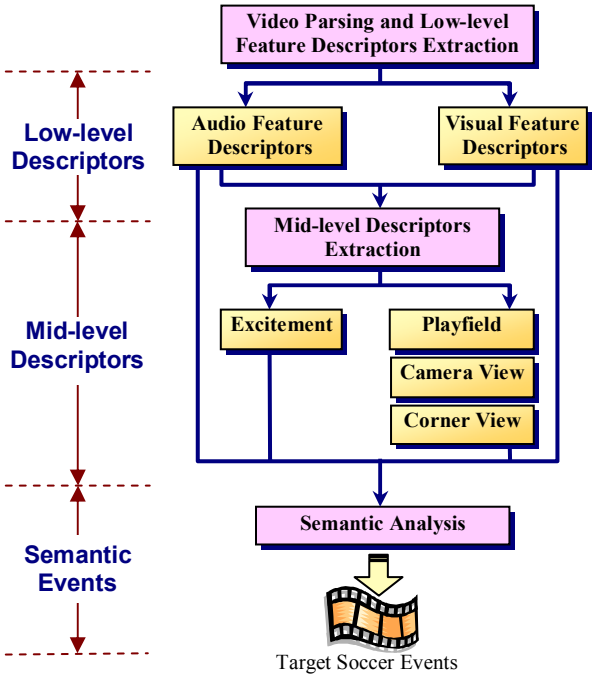
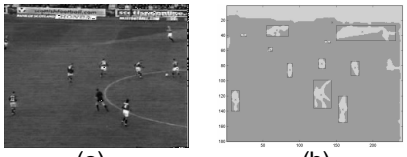


Figure 1. Framework overview

Low-level audio-visual feature descriptors can be acquired directly from the input video data in (un)compressed domain. However, due to their limited capabilities in presenting the semantic contents of the video data, it is a traditionally open problem to establish the mappings between the low-level feature descriptors and semantic events. Building mid-level descriptions is therefore considered as an effective attempt to address this problem [17]. In this paper, we introduce a group of mid-level descriptors which are deduced from low-level feature descriptors and are motivated by high-level inference (as shown in Figure 1). Such mid-level descriptors offer a reasonable tradeoff between the computational requirements and the resulting semantics. In addition, the introduction of the mid-level descriptors allows the separation of sports specific knowledge and rules from the extraction of low-level feature descriptors and offers robust and reusable representations for high-level semantic analysis using customized solutions.

2.1. Low-level processing

2.1.1. Video shot detection. Various shot-boundary detection algorithms [8][20] have been proposed in the literature. In this study, the algorithm proposed in our earlier work [3] is adopted. The basic idea is that the simpler but more sensitive checking steps (e.g., pixel-histogram comparison) are first carried out to obtain a candidate pool, which thereafter is refined by the methods that are more effective but with a relatively higher computational cost.



(a) (b)

Figure 2. An example segmentation mask map. (a) An example video frame; (b) the segmentation mask map for (a)

2.1.2. Visual feature descriptors extraction. A set of shot-level visual feature descriptors, namely *pixel_change*, *histo_change*, *class1_region_mean*, *class1_region_var*, *class2_region_mean*, and *class2_region_var* are extracted for each shot. Here, *pixel_change* denotes the average percentage of the changed pixels between the consecutive frames within a shot, and *histo_change* represents the mean value of the frame-to-frame histogram differences in a shot. In addition, the significant objects or regions of interests as well as the *segmentation mask map* of a video frame can be automatically extracted using the unsupervised object segmentation method [2]. In such a way, the pixels in each frame have been grouped into different classes (in our case, two classes called *class1_region* and *class2_region* marked with gray and white in Figure 2(b)) for region-level analysis. Intuitively, features *class1_region_mean* (*class2_region_mean*) and *class1_region_var* (*class2_region_var*) represents the mean value and standard deviation of the pixels that belong to *class1_region* (*class2_region*) for the frames in a shot. The calculation of such features is conducted in the HSI (Hue-Saturation-Intensity) color space.

2.1.3. Audio Feature Descriptors Extraction. The soundtrack of a soccer video consists of speech and vocal crowd reactions, along with other environmental sounds such as whistles and clapping. In this framework, the representations of the audio features are exploited in both time-domain and frequency-domain, which are divided into three distinct groups, namely volume related, energy related, and Spectrum Flux related features. Totally, 14 generic audio features (4 volume features, 7 energy features, and 3 Spectrum Flux features) are utilized [4].

2.2. Mid-level descriptors extraction

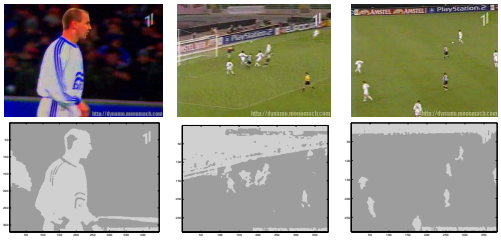
In this work, four kinds of mid-level descriptors are extracted to represent the soccer video contents.

2.2.1. Field descriptor. In sports video analysis, playfield detection is generally served as an essential step in determining other critical mid-level descriptors as well as some sport highlights. In soccer video analysis, the issue is defined as grass area detection, which remains a challenge as the grass colors may change under different lighting conditions, play fields, or shooting scales, etc. The method proposed in [7] relies on the assumption that the play field is green, which is not always true for the reasons mentioned above. The work in [6] proposed to use the dominant color based method to detect grass areas. However, its initial field color is obtained by observing only a few seconds of a soccer video. Thus its effectiveness largely depends on the assumption that the first few seconds of video are mainly field play scenes. In addition, it also assumes that there is only a single dominant color indicating the play field, which fails to accommodate variations in grass colors. In this study, an advanced strategy in grass area detection is adopted, which is conducted in three steps as given below.

Step 1: Extract possible grass areas

The first step is to distinguish the possible grass areas from the player/audience areas, which is achieved by examining the *segmentation mask maps* of a set of video frames, *S*, extracted at 50-frame interval for each shot. Compared to the non-grass areas, the grass areas tend to be much smoother in terms of color and texture distributions. Motivated by this observation, for each frame, the comparison is conducted between *class1_region_var* and *class2_region_var*, where the class with the smaller value is considered as the *background* class and its mean value and standard deviation are thus called *background_mean* and *background_var,* respectively. Correspondingly, the other class is regarded as foreground. Three sample video frames and their corresponding segmented mask maps are shown in Figure 3, where the background and foreground areas are marked with dark gray and light gray. As we can see, the grass areas tend to correspond to the background areas (see Figures 3(b) and 3(c)) due to the low variance values. For those frames with no grass area (e.g., Figure 3(a)), the background areas are much more complex and may contain crowd, sign

board, etc., which results in higher *background_var* values. Therefore, a background area is considered as a possible grass area if its *background_var* is less than T_b, which can be determined by statistical analysis of the average variation of field pixels. *grass_ratio_approx* is thus defined as the ratio of the possible grass area over the frame size. Note that the value of *grass_ratio_approx* is an approximate value, which will be utilized in step 2 to select the reference frames and will be refined in step 3.



(a) frame 1 (b) frame 2 (c) frame 3

Figure 3. (a)-(c) Three example video frames and their segmentation mask maps

Step 2: Select reference frames to learn the field colors

The reference frames are critical in learning the field colors. An ideal set of reference frames should contain a relatively high percentage of play field scenes with large grass ratios. Reference frames are selected from the shots with their *grass_ratio_approx* greater than T_{grass}. Here T_{grass} is set to the mean value of the *grass_ratio_approx* across the whole video clip. Since the feature *background_mean* represents the mean color value of each possible grass area, the color histogram is then calculated over the pool of the possible field colors collected for a single video clip. The actual play field color(s) are identified around the histogram peak(s) using the approach discussed in [4].

Step 3: Refine *grass_ratio* values

Once the play field colors are identified, the refinement of the *grass_ratio* value for a video shot is straightforward. In brief, for each segmented frame in *S*, the field pixels are detected from the background areas and thus its *grass_ratio_approx* can be refined to yield the accurate shot-level *grass_ratio* values. Note that since the background areas have been detected in step 1, the computational cost of this step is quite low. It is also worth noting that by deducing *grass_ratio* at the region-level, we resolve the problem that the non-grass areas (sign boards, player clothes, etc.) may have a close-to grass color, which was not addressed in most of the existing studies.

2.2.2. Camera view descriptor. Most of the existing camera view classification methods utilize grass ratio as an indicator of the view types, assuming that a global view (e.g., Figure 4(a)) has a much greater grass ratio value than that of a close view (e.g., Figure 4(b)) [13]. However, close view shots such as the one shown in Figure 4(c) could have large grass ratio values. Thus, the use of grass ratio alone can lead to misclassifications. In contrast, in [12], the shot view was determined via the estimation of the object size in the view. However, it is usually difficult to achieve accurate object segmentation, especially with the existence of object occlusions as shown in Figure 4 (b).

(a) Global view (b) Close view (c) Close view

Figure 4. Example camera views

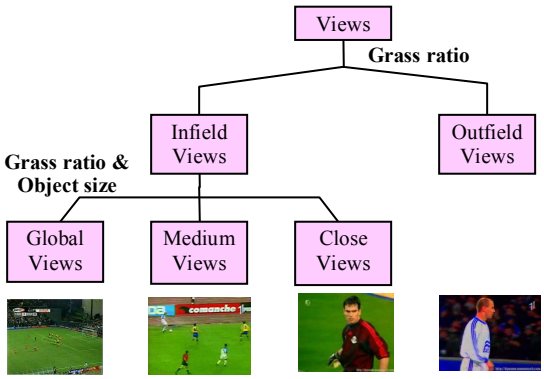


Figure 5. Hierarchical shot view

To address these issues, we propose a hierarchical shot view classification scheme as illustrated in Figure 5. Grass ratio values act as the major criterion in differentiating the outfield views and infield views. Then the infield views are further categorized into close, medium and global views using the grass ratio value coupled with the object size in the playfield. The reasons for such a setting are twofold. First, the further classification of outfield views normally fails to yield more useful information at users' interests. Thus, to simplify the problem, only the infield views are further analyzed. Second, it is relatively easier to detect the grass area as opposed to the object detection due to its homogeneous characteristic, and the proposed playfield segmentation scheme can yield quite promising results. Therefore, the grass ratio value serves as the primary differentiating factor with the facilitation of roughly estimated foreground object size in the playfield area. In brief, the foreground object with the maximal size in the field is identified, and *Max_O* is calculated to denote the ratio of its area versus the frame size. The camera view descriptor is defined in Table 1. Currently, the thresholds are

defined empirically. A statistical analysis or data classification approach might help in this manner.

Table 1. Camera view descriptors

CVD	Condition	Thresholds
Outfield	grass_ratio < T_o	T_o = 0.05
Global	grass_ratio >= T_{g1} ∧ Max_O < T_{g2}	T_{g1} = 0.4, T_{g1} = 0.05
Close	(grass_ratio < T_{c1} ∨ Max_O > T_{c2}) ∧ grass_ration > T_o	T_{c1} = 0.4, T_{c2} = 0.25, T_o = 0.05
Medium	Otherwise	

(a) Corner-kick (b) Corner-throw (c) Free-kick

Figure 6. Example corner events

2.2.3. Corner view descriptor. The corner view is defined as to have at lease one corner visible in the scene. The reason for defining the corner view lies in the fact that a large number of exciting events belong to corner events such as corner-kicks, free-kicks near the penalty box, and line throws from the corner (see examples in Figure 6). In this study, a simple yet effective approach for corner views detection is developed. The basic idea is that though the minor discrepancy or noise contained in the segmentation mask map might deteriorate the performance of the direct identification of the corner point, we can compensate and thus reduce the adverse affect of the bias by intelligently examining the size of the grass area and audience area for corner point detection.

The corner views can be generally classified into wide corner views where the corner angle is shown in the scene with no less than 90 degrees (see Figure 6(c), Figure 7(a) for examples) and narrow corner views otherwise (see Figure 6(a) and Figure 6(b)). The difference of these two classes of corner views is reflected by the relative layout and shapes of the audience areas and play fields.

Wide Corner Views

This idea is illustrated by the example in Figure 7, followed by the discussions on how to extend this idea to general wide corner view detection. Figure 7(a) shows a video frame with a wide corner view. Its corresponding segmentation mask map is given in Figure 7(b), where the black areas are the play field identified by our proposed grass area detection method and the gray areas correspond to the players, balls, and audience areas, etc. As we can see, it is quite challenging in extracting the corner point (thus the slopes of the boundary lines) automatically from the *segmentation mask map*. Therefore, we convert the problem of finding the corner point to the one of determining the location where the pattern of grass area or audience area changes dramatically. Specifically, assume the frame size is $r*c$, we define a sliding window with a fixed width d ($d << c$) to monitor the changes of non-grass area inside the window when it slides horizontally from left to right, as illustrated in Figure 7(c). Given the top-left point of the frame as the origin point, the starting position of a window n is represented by wh_n (n = 1, 2, …N), where $wh_n = d*(n-1)$ and $N = \lfloor c/d \rfloor$. Let a_n be the audience area size inside the n^{th} sliding window. When there is an observable increase of a_n (e.g., window **3** in Figure 7(c)), the corner point can be roughly located with coordinates ($d*(n-1)$, a_n/d). In our implementation, a corner point is detected when $a_n/a_{n-1}>1.05$, which is defined empirically. There are cases where the value of a_n/a_{n-1} is slightly lower than 1.05 in the presence of a corner point. According to our algorithm, the sliding window will move to the next one which causes a small shift of the actual position of that corner point. However, since $d << c$, this small shift will not affect the overall performance.

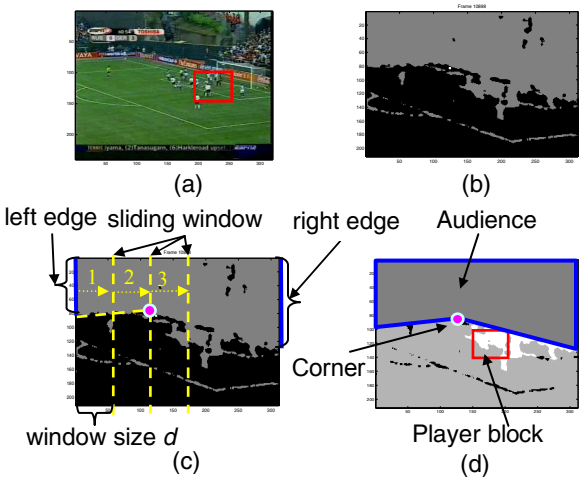


Figure 7. (a) An example frame with corner view; (b) its segmentation mask; (c) sliding window scheme; (d) the identified audience area, corner point and player block

For general wide corner view detection, two major factors should be addressed.

- The direction of the sliding window is determined dynamically by examining the spatial relationship between the grass area and audience area. Such a relationship can be roughly estimated by checking the locations of their centroid points as well as the minimal bounding boxes based on the obtained

segmentation mask map. If they are with mainly top-down (or left-right) relationship, a slide window is sliding across the frame horizontally (or vertically) starting from the edge where audience size is smaller.

- This algorithm assumes that the corner point is not contained in the first window. Therefore, window size *d* is set to a reasonable small value (10%*r or 10%*c depending on the direction of the sliding window in our study).

Note that during corner events, the majority of the players are likely to stand close to each other within a small area in front of the goal post, which forms a 'player block' with a high concentration of the player objects, as shown in Figures 7(a) and 7(d) (areas marked by red rectangle boxes). In the case of a wide corner view, such 'player block' can be detected as follows to further facilitate corner event detection. An approximate centroid of the player block is first estimate using the centroid of all the non-grass areas excluding the audience areas. Then by using another sliding window whose size is dynamically determined by the video frame size, we move it along 8 directions within a limited area around the initial centroid to locate a dense player block.

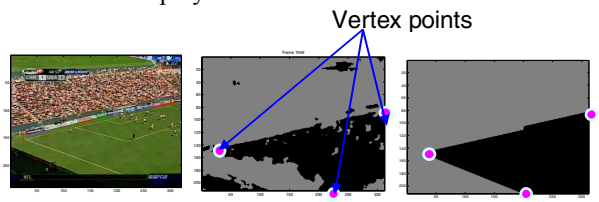


Figure 8. (a) The corner kick frame; (b) the segmentation mask map for (a) with 3 vertex points identified; and (c) the segmentation template for (b)

Narrow Corner Views

In the case of narrow corner views, the shape of the play field is close to a full triangle or a 'chopped' triangle. Figure 8 shows an example. The three vertex points of play field can be estimated by analyzing its minimum bounding box. Then a segmentation template is created based on the three vertex points as shown in Figure 8(c). If the difference between the segmentation map and its corresponding segmentation template is not significant, then this scene is regarded as a possible narrow corner view.

To generalize the idea for narrow corner view detection, we need to also determine the facing direction of the segmentation template based on the analysis of the spatial relationship between the audience and grass area.

2.2.4. Excitement descriptor. Different from the visual effects, the soundtrack of a video does not necessarily show any significant change at the shot boundary. To avoid the loss of actual semantic meanings of the audio track, the audio mid-level representation called excitement descriptor is defined to capture the excitement of the crowd and commentator in sport videos. Such an excitement is normally accompanying with or is the result of certain important events. The excitement descriptor is captured in a three-stage process. First, the audio volume feature is extracted at the clip-level. Here, an audio clip is defined with a fixed length of one second. Secondly, a clip with its volume greater than the mean volume of the entire video is extracted as an exciting clip. Finally, considering that such excitement normally last a period of time as opposed to other sparse happenings of high-volume sound (such as environmental sound) or noises, we look for a time period with multiple exciting clips to define the excitement descriptor. Here the time period is of fixed length and can be determined by adopting our previously proposed temporal pattern analysis algorithm [1]. In this study, for each shot, we examine a time period of 6 sec which includes last 3-clip portion of each shot (for short, last_por) as well as first 3-clip portion of its consecutive shot (for short, nextfirst_por). If one or more exciting clip(s) is detected in each of these 3-sec portions, we define vol_last (vol_nextfirst) to record the maximum volume of last_por (nextfirst_por) and the excitement descriptor is set to the summation of vol_last and vol_nextfirst.

3. High-level analysis

Event detection is one of the main tasks in terms of the video semantic analysis. In this section, we present a high-level semantic analysis scheme to evaluate the effectiveness of using the multimodal multi-level descriptors in event detection. Generally speaking, the semantic analysis process can be viewed as a function approximation problem, where the task is to learn a target function *f* that maps a set of feature descriptors *x* (in our case, low-level and mid-level descriptors) to one of the pre-defined event labels *y*. The target function is called a *classification model*.

In this study, the decision tree logic is used for data classification as it possesses the capability of handling both numerical and nominal attributes. In addition, it can select the representative descriptors automatically and is mathematically less complex. In brief, a decision tree is a flow-chart-like tree structure, where each internal node denotes a test on one or more

attributes (feature descriptors), each branch represents an outcome of the test, and the leaf nodes show the class distribution. In the decision tree generation process, information gain ratio criterion is used to determine the most appropriate attribute for partitioning.

As there is a wide range of events in the soccer videos, it is difficult to present extensive event detection results for all the event types. Therefore, in this study, two classes of events viz. goal events and corner events are selected for performance evaluation since they significantly differ from each other in various aspects such as event pattern and occurrence frequency. Before the decision tree based classification process starts, a feature set needs to be constructed, which contains a group of low-level descriptors, including four visual descriptors (*pixel_change*, *histo_change*, *background_mean*, *background_var*) and fourteen audio descriptors, and four mid-level descriptors. Note that since most events are the result of past activities and might cause effects in the future, to capture the temporal characteristics, these mid-level descriptors are extracted for both current shot and its two adjacent shots.

4. Experimental results

We have rigorously tested our proposed framework on a large data set with over 7 hours (432 minutes) of soccer videos, which were collected from a variety of sources, such as Euro Cup 1998, World Cup 2002, and FIFA Women's World Cup USA 2003, and are with different production/post-production styles, resolutions, and frame rates. The data set contains 3,043 video shots as parsed by the aforementioned shot detection algorithm, where the number of corner event shots and goal shots are 145 and 29, respectively.

4.1. Experimental settings

In our experiment, 2/3rds of the whole data set (called training data set) was used to train the model which was tested by the remaining 1/3rd data (called testing data set). To avoid the overfitting problem, the 5-fold cross-validation scheme is adopted for performance evaluation. The data set was randomly divided five times to obtain five different groups of training and testing data sets. Thus, five models were constructed, each tested by its corresponding testing data. Such a scheme allows better estimations of the framework's capability in applying the learned event models to other unseen data.

4.2. Event detection performance

The performance of the corner event detection is illustrated in Table 2. The "Missed' column indicates the number of false negatives, which means that the corner events are misclassified as noncorner events; whereas the 'Misiden' column indicates the number of false positives, i.e., the noncorner events being identified as corner events. Consequently, recall and precision are defined as follows:

$$\text{Recall} = \frac{\text{Identified}}{\text{Identified} + \text{Missed}}, \quad \text{Precision} = \frac{\text{Identified}}{\text{Identified} + \text{Misiden}}$$

Table 2. Performance of corner event detection

	Corner Event #	Identified	Missed	Misiden	Recall	Precision
Test 1	40	38	2	6	95.0%	86.4%
Test 2	46	45	1	9	97.8%	83.3%
Test 3	50	49	1	9	98.0%	84.5%
Test 4	43	42	1	7	97.7%	85.7%
Test 5	44	43	1	7	97.7%	86.0%
				Average	97.2%	85.2%

Table 3. Performance of goal event detection

	Goal Event #	Identified	Missed	Misiden	Recall	Precision
Test 1	11	10	1	1	91.7%	91.7%
Test 2	10	10	0	2	100.0%	83.3%
Test 3	12	11	1	2	92.3%	85.7%
Test 4	10	9	1	1	90.0%	90.0%
Test 5	11	10	1	1	90.9%	90.9%
				Average	93.0%	88.3%

As can be seen from this table, the performance is very promising, especially for the recall rate which reaches over 97% by average. In fact, in sports event detection, the metric recall is normally weighted higher than precision as we prefer to have all the targeted events detected even at the cost of including a small number of irrelevant shots. Also a further check of the experimental results finds that most misidentified shots are goal kicks/attempts whose event patterns are quite similar to that of the corner events. In fact, such events are usually considered as exciting events as well. In our future work, we will extend the current framework for goal attempts detection.

We also tested the framework upon the goal event detection and the performance is summarized in Table 3 and the results are also quite satisfactory. It should be pointed out that the goal events account for less than 1% of the total data set. The rareness of the target events usually poses additional difficulties in the process of event detection. Through the cross-

validation and multiple event detection, we demonstrate the robustness and effectiveness of our proposed framework in event detection.

5. Conclusions

In this paper, we propose a multi-level multimodal representation framework for event detection in field-sports videos. Compared with previous work in sports video domain, especially in soccer video analysis, the proposed framework is unique in its systematic way of generating, integrating, and utilizing the low- and mid-level descriptors for event detection. We argue that both low- and mid-level descriptors play important roles in event detection for sports videos, which is supported by our experimental results using a large test soccer video data set from multiple broadcast sources. Compared to the ground truth, it is shown that high event detection accuracy can be achieved. Under this framework, domain knowledge in sports video is stored in the robust multi-level multimodal descriptors, which we believe would be reusable for other field-sports videos, thus making the event detection less ad hoc.

6. Acknowledgement

For Shu-Ching Chen, this research was supported in part by NSF EIA-0220562 and HRD-0317692. For Chengcui Zhang, this research was supported in part by SBE-0245090 and the UAB ADVANCE program of the Office for the Advancement of Women in Science and Engineering. For Mei-Ling Shyu, this research was supported in part by NSF ITR (Medium) IIS-0325260.

7. References

[1] M. Chen, et al., "Semantic Event Detection via Temporal Analysis and Multimodal Data Mining," *IEEE Signal Processing Magazine*, Special Issue on Semantic Retrieval of Multimedia, 2006, vol. 23, no. 2, pp. 38-46.

[2] S.-C. Chen, et al., "Identifying Overlapped Objects for Video Indexing and Modeling in Multimedia Database Systems," *International Journal on Artificial Intelligence Tools*, 2001, vol. 10, no. 4, pp. 715-734.

[3] S.-C. Chen, et al., "Innovative Shot Boundary Detection for Video Indexing," Edited by Sagarmay Deb, *Video Data Management and Information Retrieval*, Idea Group Publishing, 2005, pp. 217-236.

[4] S.-C. Chen, et al., "A Multimodal Data Mining Framework for Soccer Goal Detection Based on Decision Tree Logic," *International Journal of Computer Applications in Technology*, in press.

[5] L.-Y. Duan, et al., "A Mid-level Representative Framework for Semantic Sports Video Analysis," *Proceedings of ACM Multimedia*, 2003, pp. 33-44.

[6] A. Ekin, et al., "Automatic Soccer Video Analysis and Summarization," *IEEE Transactions on Image Processing*, 2003, vol. 12, no. 7, pp. 796-807.

[7] Y. Gong, et al., "Automatic Parsing of TV Soccer Programs," *Proceedings of IEEE Multimedia Computing and Systems*, 1995, pp. 167-174.

[8] A. Hanjalic, "Shot-Boundary Detection: Unraveled and Resolved," *IEEE Transactions on Circuits and Systems for Video Technology*, 2002, vol. 12, no. 2, pp. 90-105.

[9] A. Hanjalic, "Generic Approach to Highlights Extraction from a Sport Video," *Proceedings of IEEE International Conference on Image Processing*, 2003, pp. 14-17.

[10] E.A. Tekalp and A.M. Tekalp, "Generic Play-Break Event Detection for Summarization and Hierarchical Sports Video Analysis," *Proceedings of IEEE International Conference on Multimedia and Expo*, 2003, pp. 169-172.

[11] D. Tjondronegoro, Y.-P. Chen, and B. Pham, "Content-based Video Indexing for Sports Applications Using Integrated Multi-modal Approach," *Proceedings of ACM Multimedia*, 2005, pp. 1035-1036.

[12] X. Tong, et al., "A Mid-level Visual Concept Generation Framework for Sports Analysis," *Proceedings of IEEE International Conference on Multimedia and Expo*, 2005, pp. 646-649.

[13] V. Tovinkere et al., "Detecting Semantic Events in Soccer Games: Towards A Complete Solution," *Proceedings of IEEE International Conference on Multimedia and Expo*, 2001, pp. 1040-1043.

[14] J. Wang, et al., "Automatic Replay Generation for Soccer Video Broadcasting," *Proceedings of ACM Multimedia*, 2004, pp. 311-314.

[15] L. Xie, et al., "Unsupervised Discovery of Multilevel Statistical Video Structures Using Hierarchical Hidden Markov Models," *Proceedings of IEEE International Conference on Multimedia and Expo*, 2003, pp. 29-32.

[16] M. Xu, et al., "Creating Audio Keywords for Event Detection in Soccer Video," *Proceedings of IEEE International Conference on Multimedia and Expo*, 2003, pp. 281-284.

[17] Q. Ye, et al., "Exciting Event Detection in Broadcast Soccer Video with Mid-level Description and Incremental Learning," *Proceedings of ACM Multimedia*, 2005, pp. 455-458.

[18] X. Yu, et al., "A Robust and Accumulator-Free Ellipse Hough Transform," *Proceedings of ACM Multimedia*, 2004, pp. 256-259.

[19] X. Yu, and D. Farin, "Current and Emerging Topics in Sports Video Processing," *Proceedings of IEEE International Conference on Multimedia and Expo*, 2005, pp. 526-529.

[20] J. Yuan, et al., "A Unified Shot Boundary Detection Framework Based on Graph Partition Model," *Proceedings of ACM Multimedia*, 2005, pp. 539-542.