Offense Based Temporal Segmentation for Event Detection in Soccer Video*

Lei Wang
Key Lab. of Pervasive Computing, Tsinghua University, Beijing, 100084, P.R.China
wanglei96@tsinghua.org.cn

Michael Lew
Leiden University, Niels Bohrweg 1, 2333 CA Leiden, Netherlands
mlew@liacs.nl

Guangyou Xu
Key Lab. of Pervasive Computing, Tsinghua University, Beijing, 100084, P.R.China
dcs-xgy@tsinghua.edu.cn

ABSTRACT

Sports video is regarded as a good testing bed for techniques on content based video analysis and processing. Although partially successful systems have been designed for specific sports domains with limited data, most previous works do not adequately address the problem of temporal segmentation for event detection, nor the event representation problem. In this paper, we present an analysis of soccer video for detecting the semantic notion of *offense*. It is not only useful as a new semantic concept of sports video analysis, but also provides temporal segmentation for video event detection and representation. We propose a system to detect the offensive unit in soccer video automatically. The offensive unit is then used to calculate new semantics like *possession*, as well as to detect goal events in video. Experimental results on various sources of soccer video have verified that our approach extracts the new semantic notions successfully and facilitates video event detection and representation.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video analysis*; I.4 [**Image Processing And Computer Vision**]: Scene Analysis

General Terms

Algorithms, Design, Experimentation

Keywords

events, temporal segmentation, semantics, sports video

1. INTRODUCTION

Sports video is regarded as a good testing bed for techniques on content based video analysis and processing. It involves a variety of problems such as video indexing, semantic analysis, video retrieval, video summarization and streaming. Concepts and events are generally well defined in sports video, and sports video applications are strongly driven by the popularity of different sports, such as soccer, basketball and tennis.

In the literature, previous works have focused on several key aspects. One of these is extracting different cues for sports video analysis, such as slow-motion replays [16, 10, 15], dominant colors (playground colors) [5, 12], player uniform colors [21], ball trajectories [25, 7, 17, 9, 19, 26], goal mouths [25, 20] and penalty areas [6], captions [13, 14] and texts [2], and audios [18, 8]. Another primary area of investigation is the segmentation of sports video into different categories (mostly based on shots), such as play/pause detection [5, 11] and structure analysis [7, 24, 22]. A third area is the extraction of different semantic levels and applications, such as sports video summarization [6], highlights [25, 8], and detection of specific high-level events like goals [5], shots [7], and other activity [23].

Detection of high-level sports video events is particularly challenging because current event detection techniques suffer from the difficulties of temporal segmentation, i.e., determining which features in what time period should be used for event detection or reasoning. Further problems are presented by event representation, such as what video segment of a detected event should be presented to the user for a more complete experience. These problems are not well addressed in most previous works, in which a shot or a predefined length of video were taken as the temporal unit for event analysis [3]. Some works [5, 11] used a slip window or fixed window around some exciting occurrences such as slow-motion replay segments for event detection and a fixed window length for event representation, such as 15 seconds or 30 seconds. Other HMM based techniques [23] working at the frame level tend to avoid the temporal segmentation problem by determining video states and transitions of states simultaneously. However, as it is well known, this approach is less stable and highly dependent on the training data and test data.

A suitable temporal segmentation should: 1) reflect the inherent structure of video data; 2) coincide with human understanding and facilitate user experience; 3) be practical from the technical point of view. In this paper, we present a semantic notion of "offense" as the temporal unit for sports video analysis. It is defined as a complete attempt of a team (player) in an opponent sport to score a point.

The notion of offense exists in most opponent sports, es-

*This research is supported by Chinese NSF grant No. 60273005.

MIR'04, October 15–16, 2004, New York, New York, USA.

pecially popular sports such as soccer, basketball, volleyball, football and tennis. Offense itself has very clear semantics, and provides very interesting information for video summarization and highlight extraction. A change in the offensive team is moreover a natural temporal segmentation criterion for event detection and representation, and is practical and possible to be extracted automatically and efficiently.

A system for offense segmentation and offense based event detection in soccer video is presented in this paper. The soccer video is first parsed into shots, and the shots are classified based on automatically extracted playground color and player uniform colors. Camera motion of each frame is then estimated as an indicator of offense direction in global view shots, and adjacent frames with the same offense direction are grouped into the same offense segment together with non-global view shots that follow them. Domain knowledge is applied to further refine the offense segmentation and to handle problems like backward passing. Once the offense level is established, possession percentages could be calculated by summing up the offense time of each team, and the goal event is detected with multiple extractable cues within each offense.

The major contributions of our work in this paper include the following:

1. a novel semantic notion of "offense", which is the unit for our video analysis and event detection;
2. a complete novel system to robustly extract "offense" from video data automatically, and in addition, to calculate "possession";
3. a novel offense based Bayesian network for soccer video event (goal) detection.

The paper is organized as follows: Section 2 describes the definition of offense, which is our temporal unit for sports video analysis. Our system for soccer video analysis and event detection based on the semantic notion of offense is introduced in Section 3, with experimental results presented in Section 4. We discuss future extensions in Section 5 and conclude this paper in Section 6.

2. OFFENSE AS THE TEMPORAL UNIT OF ANALYSIS

Offense is a general semantic notion of sports video, especially for opponent sports like soccer, basketball, volleyball, football, and tennis. It is defined as a complete attempt to score a point. It could be a serve in tennis, a hit in table tennis, or a possession of the ball in soccer or basketball, for example. Note that a "complete" offense is not necessarily a "successful" offense, and offensive actions may include some seemingly defensive activities such as a block in volleyball, since could possibly lead to a point. And sometimes, offense may produce a point for the opponent, e.g., smashing the ball out of bounds. In other sports like soccer and basketball, although offense is an attempt for a point, it may lead to an alternate result such as another offensive attempt (from an opponent's foul).

Obviously, not all offense leads to a point, but each point is based on some offense. As a result, offense is a natural temporal segmentation criterion for event detection. In other words, division of video into segments of offense facilitates detection of events within each corresponding period. In addition, since offense is a *complete* attempt to score a point, it is usually appropriate to present the offense segment with the detected event for a complete user experience for the event. This is nominally better than a fixed window which may include some irrelevant video or miss some important context of the event.

Offense is not only a temporal segmentation criterion for event detection and representation. It also provides very useful information for video summarization and highlight extraction. For example, in soccer video, offense time and frequency are important components of possession percentage. Segmentation of offense can also help coaches in analyzing players and team behavior as the score changes. With temporal and motion analysis for the offense, it is also very easy to find fast breaks in soccer or basketball video automatically. In tennis and volleyball videos, while points with only one offense are obviously serve-score, points with numerous occurrences of offense (for both sides) are often the most exciting segments of the game.

In sports video like tennis and volleyball, there are usually some fixed cameras shooting the whole playing area. Offense could be detected by tracking and analyzing the ball trajectory in the video with respect to the net and the boundary lines [17]. For soccer and basketball videos, the ball can be difficult to track, but when the ball trajectory is extractable [26] it is useful for offense analysis. To improve robustness, we propose to detect offense by analyzing the motion of players, which is usually consistent with the camera motion, since the camera is always trying to catch the ball. We will further explain this technique in the following section.

It is interesting to note that, different from most previous works, a shot is not taken as the unit for video analysis in our work. As has been mentioned in previous works [11], a shot in sports video is not a unit for telling a story. It has little semantic meaning in sports video. Several offensive series may occur within just one shot, and a shot could also be just a replay of previous events. Even though the shots are grouped into in play segments and out-of-play segments [3], a shot based segment would be too long for an event like a goal if only the last third of the global view shot is really relevant to the goal. So a shot is not an ideal segmentation indicator for event detection or hierarchical video browsing and retrieving. However, as we show in the following section, different shot types do convey very important information about the game. As a result, shot segmentation and classification is still beneficial in sports video analysis and event detection.

The notion of offense is designed for general usage in sports video, and as discussed in this section, offense analysis is applicable to a broad range of sports, especially popular sports. Although offense may be too minute as a unit for sports like tennis and volleyball, other context based temporal segmentation, such as a score, is applicable. The importance of a context based temporal segmentation in an event detection framework is what we want to highlight here. In this paper, we demonstrate its application in video analysis and event detection for the sport of soccer, although these methods could be extended to other sports.

3. SYSTEM FOR OFFENSE BASED SOCCER VIDEO ANALYSIS

We developed a system to verify and demonstrate the ap-

plication of offense in sports video analysis. Currently the system is designed for soccer video only, and we are working on other sports types.

An overview of the system is shown in Figure 1. To achieve the offense segmentation of a soccer video, shots are segmented and classified into global view, close view and out-of-field. Camera motion is estimated to initialize the offense parsing, while domain knowledge is utilized to further refine the segmentation. Once the video is parsed into offenses, a new temporal segmentation is established for event detection. Events could be detected and represented based on each offense. Essential techniques for event inference with multiple cues are developed as well. The technical details of our algorithms are presented in the following sections.

3.1 Shot Segmentation

The raw video data is first segmented to shots by calculating the motion-compensated block-based frame difference, which is defined as

$$D = \sum_{B_k \in G} (\min_{(u,v)} \sum_{(x,y) \in B_k} (I_j(x,y) - I_{j-1}(x+u, y+v)))$$

where G is the set of all blocks (block size: 16×16), B_k is a block, $I_j(x,y)$ is the color of pixel (x,y) in frame j, and (u,v) is within a search range of the motion vector. This method avoids the problem of high color correlation in sports video that exists for color histogram based shot segmentation, and it is robust to local variances like camera and object motions.

3.2 Semantic Color Extraction

Color is a widely used feature for shot classification [6, 24, 4]. In [21], an automatic algorithm is proposed to estimate the dominant color and the player uniform colors with Gaussian Mixture Models. The dominant color is usually the playground color in sports video. The model is initialized by two separate peaks in the color distribution and then estimated with the EM algorithm. The first peak is the overall histogram peak, while the second peak color is the most frequent color that is reasonably distinct from the first peak color. The extracted color model is tested against whether it is actually a single color, based on the extracted model parameters, and is further refined based on domain constraints of sports videos. The player uniform colors are extracted automatically as well by locating body regions using face detection results. We refer the readers to [21] for technical details.

3.3 Shot Classification

Semantic color extraction provides very useful features such as the playground color ratio R_g and player uniform color ratio R_p (which are ratios between the number of corresponding color pixels and the number of pixels in a frame) for shot classification. Note that R_p takes into account the color pixels of both teams. Morphological operations (three erosion and three dilation operations in our system) are performed to eliminate the noise. In our system, all the shots are classified into three categories: global view, close view (close up and medium distance in the field), and out of field.
Global View: $R_g > T_g$ and $R_p \leq T_p$ (a large playground color ratio, and a small player uniform color ratio)
Out-of-field: $R_g \leq T_g$ and $R_p \leq T_p$ (a small playground color ratio, and a small player uniform color ratio)
Close View: $R_p > T_p$ (a large player uniform color ratio)

where T_g and T_p are two thresholds trained off-line by SVM with ten minutes of labelled data. This approach performs much better results than current techniques. Other previous works [24, 6] have determined more categories, particularly close up and medium views. However, according to our observations on video data, the many types of close up and medium shots make it difficult to accurately distinguish them, not only by automatic algorithms but also by humans. These two types of views usually convey very similar semantics and game information anyhow. For example, both close up views and medium views are commonly used to track key players.

3.4 Offense Detection

The main idea of offense detection is to segment adjacent frames in a global view with consistent camera motion into individual offense attempt, where all non-global views (including slow-motion replays) are incorporated with the preceding offense. As we have discussed, the camera motion is usually consistent with the offense direction.

3.4.1 Camera Motion Estimation

There are several reasons for choosing only global views to estimate camera motion and determine offense:

1. Most global cameras for a soccer game are put on the same side of the playground to avoid confusing the audience. Consequently, the camera motion of global views is consistent with the offense direction, while the camera motion of closer views is arbitrary;

2. Camera motion estimation is more accurate for global views and is generally insensitive to object motions in comparison to close up and medium views, which usually contain significant object motions;

3. In sports video, most close up and medium views highlight previous events or key players in these events, and are often shown during game breaks [24]. It is therefore reasonable to combine them with the previous global view of an offense.

Camera motion of each frame in global view shots is extracted by block-based motion estimation (block size: 16×16) as used in MPEG encoders, which can potentially allow real-time processing. Adjacent frames of the same camera motion direction could be categorized into the same offense segment. Here, only the horizontal component of a motion vector m_x is considered, and the offenses are correspondingly classified into two classes: Left-to-Right and Right-to-Left. m_x is smoothed to avoid noise in motion estimation. Figure 2 (a) shows the smoothed camera horizontal motion for a sample video (World Cup A1 frames 0 to 4000).

3.4.2 Initial Offense Segmentation

Table 1 shows the pseudo-code and the algorithm description for initial offense segmentation. All the non-global view shots are grouped into the previous offense. In global view shots, sometimes the camera is nearly stationary, or zooming into the field, thus the m_x is very close to zero and the corresponding frames are parsed into the previous offense. When m_x is large enough for a global view frame, two possible cases exist: 1) the direction of m_x is consistent with its previous offense, and obviously the frame should be parsed into the previous offense; 2) the direction of m_x is different

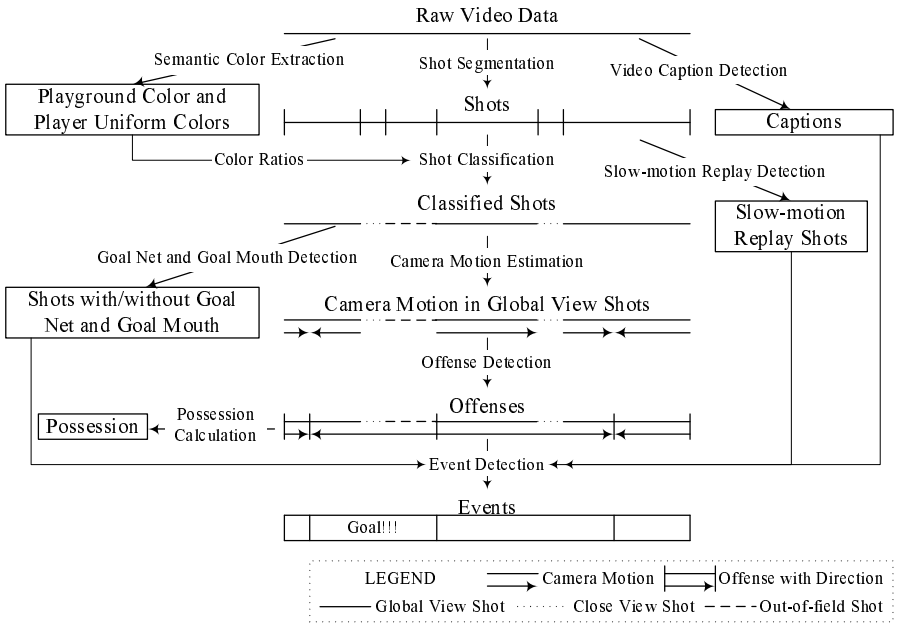

Figure 1: The overview of our system.

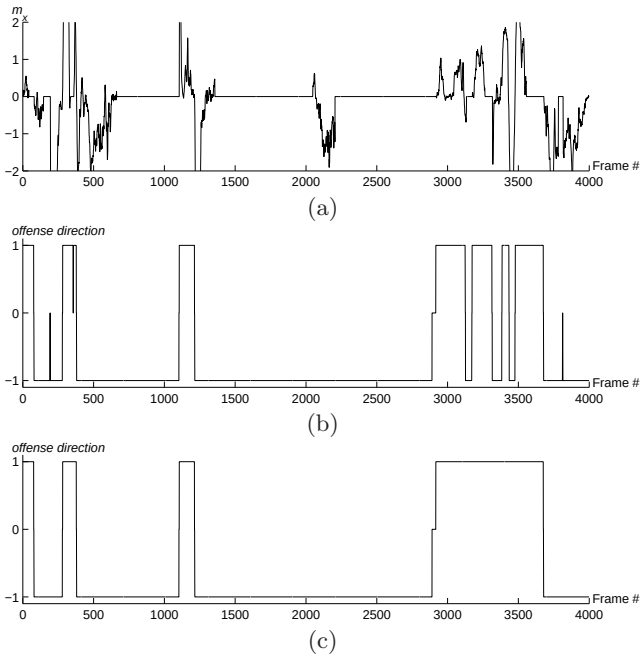

Figure 2: Offense detection based on shot classification and camera global motion estimation.

from its previous offense or there is no offense, which indicates a new offense in the direction of m_x. Since most global view cameras are placed on the same side of the stadium to avoid confusing the TV audience, the meaning of camera motion is consistent in the game. Figure 2 (b) shows the initial offense segmentation for the frame motions shown in Figure 2 (a).

3.4.3 *Offense Merging*

Different offensive behaviour in soccer videos should be addressed for better offense segmentation. The first one is backward passing. As shown in Figure 2 (b), there are frequent switches in offense from frame 2800 to 3700, which is actually fore-and-aft passing of a team, quite common in soccer video. Taking into account that a meaningful offense should last for some time, it is reasonable to combine those extremely short offenses for a smooth visual experience of the users browsing the video clip.

In addition, two consecutive offenses in the same direction and of very short intervals (e.g., around frame 200, 300 and 3800 in Figure 2 (b)) should be combined. These divisions were caused by the initial offense segmentation algorithm which concludes an offense when another global view appears. In soccer video programs, it is not uncommon for some close views of the player controlling the ball to be inserted into an offensive segment mostly with a global view, but these close views should not separate the offense.

Furthermore, when the offense reaches the penalty area but fails to score, usually the close view shots and replay shots are shown only after the offense fails (due to an offensive foul or turnover). However, a turnover leads to a change in offense. According to the initial offense segmentation algorithm, this would attach the non-global views to the very short offense of a single turnover kick. This problem should be tackled as well.

Based on these considerations, we further merge and refine the initial segmented offenses using the following domain related rules (sorted by priority):

1. An offense with a slow-motion replay shot or no fewer than two non-global view shots at the end is *never* merged with the subsequent offense;

2. If a very short offense ($L < T_o$, where L is the length of the shot and T_o is a threshold) is just between two offenses with the same direction, these three offenses

```
offcrtv = false;   // whether an offense has been created in the video
offcrts = false;   // whether an offense has been created in the shot
for (all shots) {
  if (NOT global view) {   // current shot is non-global view
    if (offcrtv) {   // offense has been created in the video
      append the shot to current offense;
  } }   // else, noop;
  else {   // current shot is global view
    if (offcrts), end current offense;   // in this step, each offense has only one global view part.
                                         // an offense ends when meets a new global view shot
    offcrts = false;   // no offense has been created in the shot
    for (all frames in the shot) {
      if (abs(m_x) <= threshold) {   // no significant horizontal motion
        if (offcrts), append current frame to current offense;   // offense has been created in the shot
      }   //else, noop;
      else {   // significant horizontal motion
        if (offcrts) {   // offense has been created in the shot
          if (m_x * od > 0) {   // frame camera motion is consistent to current offense direction
            append current frame to current offense;
          }
          else {   // frame camera motion is not consistent to current offense direction
            end current offense;
            create a new offense, od = m_x;
        } }
        else {   // no offense has been created in the shot
          create a new offense, od = m_x;
          offcrts = offcrtv = true;   // an offense has been created in the video and the shot
} } } } }
```

Table 1: Pseudo-code for initial offense segmentation.

are merged into one ignoring the direction of the short offense;

3. If a very short offense ($L < T_o$) is between two offenses with different directions, merge it with the adjacent offense that shares the same direction;
4. If two adjacent offenses have the same direction, and the previous one has a single short ($L < T_o$) close view shot at the end, merge these two offenses;
5. If two adjacent offenses have different directions, and the latter one has a short ($L < T_o$) global view segment attached to close view shots or slow-motion replay shots, these two offenses are merged into one offense with the direction of the first offense.

These rules are depicted in Figure 3 as well. In our system, T_o is set to 2 seconds for all the experiments, based on user study.

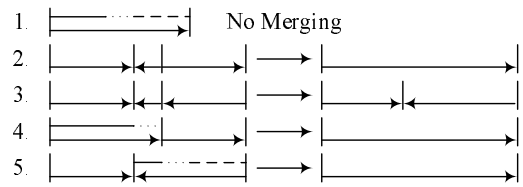

Figure 3: Offense merging handles different offensive behaviour in soccer video.

The results for the sample video clip after merging are shown in Figure 2 (c), problems in the initial offense segmentation have been solved well. For example, most separate backward passes are parsed into the correct offense. Experimental results also show that even if a series of backward passes is misclassified as an offense segment for the opposing team, such a misclassification is not serious, since it usually does not contain important plays that would be shown to the user.

3.5 Possession Calculation

Possession percentage is often correlated to how much a team is controlling a game. With the segmented offense representation, one can easily count the possession percentage by summing the offense time for each team and calculating the ratios. Here, only the time of global view frames are calculated, since they usually correspond to game play, while non-global view frames are regarded as breaks in play.

3.6 Goal Detection

Bayesian networks are widely used for multi-modal feature fusion and event detection [5]. The major difference between our work and previous works are: 1) the event is detected and presented based on each offense, while previous works usually detect events in a fixed temporal window, like 15 or 30 seconds around some exciting occurrence; 2) several new and useful cues are employed in our system.

The cues we used for the Bayesian network include:

1. C_1: the length of the offense t_o;
2. C_2: the length of game pause, e.g., the length of non-global view shots t_n;
3. C_3: the number of close view shots n_c;
4. C_4: the number of slow-motion replay shots n_s;

5. C_5: the number of out-of-field shots n_o;

6. C_6: the number of shots containing the goal net n_{ne};

7. C_7: the number of shots containing the goal mouth n_g;

8. C_8: whether a caption is shown in the next offense b_c;

C_1, C_2, C_3 and C_5 are directly extractable from the offense segmentation and shot classification. The slow-motion replay shots (C_4) are detected based on [10, 16]. Only the goal mouth (C_7) is detected in global view shots using the technique described in [20]. And video captions (C_8) are detected by analyzing the statistical pixel changes in the sequence [14, 1]. A detailed discussion of these existing techniques is beyond the scope of this paper, and we refer the readers to the references above.

The goal net (C_6) is only detected in the close view shots by texture analysis. A 17-D vector is used to represent a region, and a Gaussian Mixture Model for net detection is trained based on off-line labelled data. In the detection phase, the frame of a close view shot is subsampled at different scales. At each scale, the subsampled image is divided into blocks (16×16 in our system), and each block is represented by the 17-D texture descriptor and then classified into "net" or "non-net" according to the model. If more than one third of the blocks are classified as "net" at some scale, the frame, and subsequently the shot, is regarded as having a net in it.

The goal event is denoted by X. So the problem is to solve

$$X = \arg\max_X P(X|C_1, C_2, \cdots, C_8)$$

as the cues are detected. The states of the stochastic variables X and C_i are shown in Table 2. In our system, like previous works, we assume that all the C_is are conditionally independent with respect to X. Although a full connected Bayesian network would generate better results, it is very difficult to find enough data to train such a complex model, especially for the goal state, e.g., $X = 1$, since goals occur seldomly in soccer video. As a result, many items like $P(C_3|X = 1, C_1, C_2)$ in a fully connected Bayesian network would be zero while other items may be 1. This definitely reduces the ability to expand the model. So we suggest utilizing a simple Bayesian network based on conditional independence of all the cues, and try to extend to a more complex model when we have more training data, especially more offenses with goals, from the retrieval process using the user feedback. Now the problem becomes

$$\begin{aligned} X &= \arg\max_X P(X|C_1, C_2, \cdots, C_8) \\ &\propto \arg\max_X P(X)P(C_1|X)P(C_2|X)\cdots P(C_8|X) \end{aligned}$$

The model is established with training data. And once the cues are detected, the goal event could be inferred by solving this equation.

4. EXPERIMENTS

We tested our system with 200 minutes of soccer video from various sources: three from World Cup 2002 (WC), two from the English Premier League 2003 (EPL), and one from the German Soccer League - Bundesliga 2003 (GSL). No training data was used for shot segmentation and offense segmentation. The thresholds for shot classification are extracted based on a ten minutes training data and a SVM classifier. And the model for goal detection is trained on ten other soccer video clips with add up to more than 5 hours.

4.1 Shot Segmentation and Classification

The experimental results for shot segmentation using our motion-compensated block-based frame difference with shot classification from playground color ratio and player uniform color ratios are shown in Table 3 and Table 4 separately.

4.2 Offense Detection

The experimental results for offense detection based on motion analysis and merging based refinement are shown in Table 5. Since the offense itself is a semantic concept of sports video, unlike shot segmentation, there is no ground truth for offense detection results. As a result, soccer fans who are unaware of our research work are asked to evaluate the results. The evaluation has three levels: correct, nondescript, and false. The "nondescript" level is applied to some game segments where the two teams are fighting to control the ball. It is difficult to distinguish the offense direction in this case. Incorrect segmentation is mainly due to a series of backward passes, which is misclassified as an offense segment for the opposing team. However, critical events rarely occur in these segments, so actually they make no difference to the results (especially for event detection).

4.3 Possession Calculation

We compared our results for possession percentage based on the offense time with officially released data, which is shown after each half of the soccer game. However, only four video clips have this figure shown in the video. The results are shown in Table 6, which are very close to official data. The differences could be attributed to two reasons: 1) some close up or medium views of play are not counted into the possession time in our approach; 2) when a team A passes the ball backward frequently, it could be counted as offense time for the opposing team B. It is reasonable from some perspective, since although the opposite team B is not controlling the ball, it controls the game actually as it presses the offense of team A. It should also be considered that even for official data, different TV channels and sports analysts may have slightly different results.

4.4 Goal Detection

Table 7 shows the results for goal detection in our framework. As can be seen in the table, most goals have been detected correctly. According to our further investigation, the missed goals are due to quick starts in the mid-field after the goal and delayed slow-motion replays. And falsely detected goals are attributed to some goal reviews during game pauses and some exciting shots that seem to be goals.

For comparison, we also tried fixed slip window based temporal segmentation for goal detection. However, we found that the evaluation is difficult to make since the results are highly dependent on the length of the window. A goal segment in different matches lasts from twenty seconds to more than one minute. As a result, even though the fixed window based approach detects a goal, it often does not present the event appropriately to the user. In addition, the slip window based calculation is more time consuming.

We think that exciting goal area events are more important than less exciting segments to the users. In other words,

	0	1	2	3	4	5
X_i	no goal	goal	-	-	-	-
C_1	$t_o \in [0, 5)$	$t_o \in [5, 10)$	$t_o \in [10, 20)$	$t_o \in [20, 40)$	$t_o \in [40, 80)$	$t_o \in [80, \infty)$
C_2	$t_n \in [0, 1)$	$t_n \in [1, 2)$	$t_n \in [2, 6)$	$t_n \in [6, 18)$	$t_n \in [18, 60)$	$t_n \in [60, \infty)$
C_3	$n_c = 0$	$n_c = 1$	$n_c = 2$	$n_c = 3$	$n_c \in [4, 6)$	$n_c \in [6, \infty)$
C_4	$n_s = 0$	$n_s = 1$	$n_s = 2$	$n_s = 3$	$n_s = 4$	$n_s \in [5, \infty)$
C_5	$n_o = 0$	$n_o = 1$	$n_o = 2$	$n_o = 3$	$n_o = 4$	$n_o \in [5, \infty)$
C_6	$n_{ne} = 0$	$n_{ne} = 1$	$n_{ne} = 2$	$n_{ne} = 3$	$n_{ne} = 4$	$n_{ne} \in [5, \infty)$
C_7	$n_g = 0$	$n_g = 1$	$n_g = 2$	$n_g = 3$	$n_g = 4$	$n_g \in [5, \infty)$
C_8	$b_c = false$	$b_c = true$	-	-	-	-

Table 2: The state of the stochastic variables and their corresponding physical meaning.

Sequence	WC A1	WC A2	EPL B1	EPL B2
Shots (ground truth)	314	292	232	242
Shots (correctly detected)	295	277	216	225
Detection rate (%)	93.8	94.7	93.2	92.9

Table 3: Experimental results for shot segmentation.

Sequence		WC A1			WC A2			EPL B1			EPL B2	
	G.	C.	O.	G.	C.	O.	G.	C.	O.	G.	C.	O.
ground truth	125	162	27	112	160	20	84	129	19	87	130	25
G. (detected)	122	4	0	110	1	1	82	2	0	86	3	0
C. (detected)	3	146	7	2	152	4	2	119	6	1	116	9
O. (detected)	0	12	20	0	7	15	0	8	13	0	11	16

Table 4: Experimental results for shot classification. (G. - Global view, C. - Close up and medium, O. - Out of field)

Sequence	WC A1	WC A2	WC A3	EPL B1	EPL B2	GSL C1
Offenses (ground truth)	159	142	140	94	98	180
Offenses (correctly detected)	138	121	124	82	89	168
Offenses (nondescript)	5	4	3	3	3	4
Offenses (falsely detected)	16	17	13	9	6	8
Accuracy (%)	86.8	85.2	88.6	87.2	90.8	93.3

Table 5: Experimental results for offense detection.

Sequence	WC A1	WC A2	EPL B1	GSL C1
Possession (official release, %)	49:51	55:45	58:42	50:50
Possession (calculated, %)	48.7:51.3	55.5:44.5	56.8:43.2	49.3:50.7

Table 6: Experimental results for possession calculation.

Sequence	WC A1	WC A2	WC A3	EPL B1	EPL B2	GSL C1
Goals (ground truth)	2	3	1	0	5	1
Goals (correctly detected)	1	3	1	0	4	1
Goals (missed)	1	0	0	0	1	0
Goals (falsely detected)	0	1	1	2	1	0

Table 7: Experimental results for goal detection.

the users prefer to view a segment that is not a goal but a good scoring opportunity (i.e., goal falsely detected), and also not miss actual goals. In news highlights, these exciting but unsuccessful scoring opportunities are often shown. So our system is designed to trade off some accuracy for a better detection rate. For example, some states are nearly impossible for the offenses with goals, such as when the offense has only one slow-motion replay shot, e.g., the data shows that $P(C_4 = 1|X = 1) = 0$. However, in order to avoid missing a goal which is very strong in the other cues but weak in C_4, we still set $P(C_4 = 1|X = 1)$ to a non-zero small possibility. As can be seen in the results, this consideration is reasonable. With just several false detections, our system extracts nearly all the goals in the video. Compared to other reported results [5, 3] that utilize different cues, our results are better with less false alarms and a more reasonable temporal representation for the events.

5. DISCUSSIONS

The most time consuming part of our system is the block-based camera motion estimation for each frame (the block-based frame difference for shot segmentation is extracted in this process too). However, our experiments show that the speed is still in real time. On a standard PC with Athlon 2500+ CPU and 512M RAM, using the block-based matching algorithm of OpenCV and then calculating the camera motion, the average time for processing a 352×288 image is 0.01577 seconds, which is equivalent to 63 frames per second, clearly real time. And this time includes decoding the DivX compressed AVI. If the motion vectors are extracted directly from the MPEG file, the speed could be even faster.

Although our system is able to run in real time, in the whole process, we need to iteratively utilize the context of the video data to refine the results (such as semantic color extraction). The context is also essential for event detection, since the slow-motion replay shots and captions are all very important cues that appear after the event has occured. Consequently, our system takes a whole video as the processing unit.

While several techniques such as shot segmentation and shot classification are really common in sports video analysis [1], what we really want to emphasize in this paper are context based temporal segmentation for video analysis and event detection based on multiple cues with domain knowledge. Although different temporal segmentation units and cues would (and should) be employed in different sports, the concept is very similar. We are currently examining how to apply this framework to different sports videos, such as volleyball and tennis.

6. CONCLUSIONS

Temporal segmentation is a critical problem for video event detection, which we have addressed in this paper. A new semantic notion of *offense* is proposed accordingly, which is not only a useful new semantic concept for sports video analysis, but also an excellent temporal segmentation cue for video event detection. A system is described to automatically detect the offense in soccer video based on shot classification and motion analysis. Each offense segment is then used to calculate new semantics like possession, as well as to detect goal events in video. Experimental results on various sources of soccer video have verified that our approach extracts the new semantic notions successfully and facilitates video event detection.

Acknowledgment

The authors would like to thank Steve Lin of Microsoft Research Asia for proofreading the paper; Guoying Jin and Linmi Tao of Tsinghua University and Ying Pan of Peking University for fruitful discussions and their help in preparing the experimental data.

7. REFERENCES

[1] J. Assfalg, M. Bertini, C. Colombo, and A.D. Bimbo. Semantic annotation of sports videos. *IEEE Multimedia*, 9(2):52–60, April/June 2002.

[2] N. Babaguchi, Y. Kawai, and T.Kitahashi. Event based video indexing by intermodal collaboration. *IEEE Transactions on Multimedia*, 4(1):68–75, March 2002.

[3] L.-Y. Duan, M. Xu, T.-S. Chua, Q. Tian, and C.-S. Xu. A mid-level representation framework for semantic sports video analysis. In *ACM Multimedia*, pages 33–44, 2003.

[4] L.-Y. Duan, M. Xu, Q. Tian, and C.-S. Xu. Nonparametric color characterization using mean shift. In *ACM Multimedia*, pages 243–246, Berkeley, CA, USA, 2003.

[5] A. Ekin and A.M. Tekalp. Generic event detection in sports video using cinematic features. In *Second IEEE Workshop on Event Mining (EVENT'03)*, pages 17–24, June 2003.

[6] A. Ekin, A.M. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. *IEEE Transactions on Image Processing*, 12(7):796–807, July 2003.

[7] Y. Gong, L.T. Sin, C.H. Chuan, H. Zhang, and M. Sakauchi. Automatic parsing of tv soccer programs. In *IEEE International Conference on Multimedia Computing and Systems (ICMCS'95)*, pages 167–174, 1995.

[8] A. Hanjalic. Generic approach to highlight detection in a sport video. In *ICIP'03*, volume 1, pages 1–4, 2003.

[9] S. S. Intille and A. F. Bobick. Recognizing planned, multi-person action. *CVIU*, 81(3):414–445, March 2001.

[10] V. Kobla, D. DeMenthon, and D. Doermann. Identification of sports videos using replay, text, and camera motion features. In *SPIE Conference on Storage and Retrieval for Media Databases*, volume 3972, pages 332–343, 2000.

[11] B.X. Li, H. Pan, and M.I. Sezan. A general framework for sports video summarization with its application to soccer. In *ICASSP'03*, volume 3, pages 169–172, April 2003.

[12] B.X. Li and M.I. Sezan. Event detection and summarization in american football broadcast video. In *SPIE Conference on Storage and Retrieval for Image and Video Databases*, volume 4676, pages 202–213, 2002.

[13] H. Li and D. Doermann. Automatic identification of text in digital video key frames. In *ICPR'98*, 1:129–132, 1998.

[14] B. Luo, X. Tang, J. Liu, and H.-J. Zhang. Video caption detection and extraction using temporal information. In *ICIP'03*, pages 297–300, Barcelona, Spain, September 2003.

[15] H. Pan, B.X. Li, and M.I. Sezan. Automatic detection of replay segments in broadcast sports programs by detection of logos in scene transitions. In *ICASSP'02*, 4:3385–3388.

[16] H. Pan, P. van Beek, and M.I. Sezan. Detection of slow-motion replay segments in sports video for highlights generation. In *ICASSP*, volume 3, pages 1649–1652, 2001.

[17] G.S. Pingali, Y. Jean, and I. Carlbom. Real time tracking for enhanced tennis broadcasts. In *CVPR'98*, 260–265.

[18] Y. Rui, A. Gupta, and A. Acero. Automatically extracting highlights for TV baseball programs. In *ACM Multimedia*, pages 105–115, Los Angeles, CA, 2000.

[19] V. Tovinkere and R.J. Qian. Detecting semantic events in soccer games: Towards a complete solution. In *ICME'01*.

[20] K.W. Wan, X. Yan, X. Yu, and C. Xu. Real-time goal-mouth detection in mpeg soccer video. In *ACM Multimedia*, pages 311–314, Berkeley, CA, USA, 2003.

[21] L. Wang, B. Zeng, S. Lin, G. Xu, and H. Shum. Automatic extraction of semantic colors in sports video. In *ICASSP'04*, Montreal, Canada, May 2004.

[22] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun. Structure analysis of soccer video with hidden markov models. In *ICASSP'02*, volume 4, pages 4096–4099, May 2002.

[23] G. Xu, Y.-F. Ma, H.-J. Zhang, and S.Q. Yang. A HMM based semantic analysis framework for sports game event detection. In *ICIP'03*, pages 25–28, September 2003.

[24] P. Xu, L. Xie, S.F. Chang, A. Divakaran, A. Vetro, and H. Sun. Algorithms and systems for segmentation and structure analysis in soccer video. In *ICME'01*, 928–931.

[25] D. Yow, B.L.Yeo, M. Yeung, and G. Liu. Analysis and presentation of soccer highlights from digital video. In *ACCV'95*, pages 499–503, Singapore, December 1995.

[26] X. Yu, C. Xu, H.W. Leong, Q. Tian, Q. Tang, and K.W. Wan. Trajectory-based ball detection and tracking with applications to semantic analysis of broadcast soccer video. In *ACM Multimedia*, pages 11–20, Berkeley, CA, 2003.