Fusion of AV Features and External Information Sources for Event Detection in Team Sports Video

HUAXIN XU and TAT-SENG CHUA
National University of Singapore

The use of AV features alone is insufficient to induce high-level semantics. This article proposes a framework that utilizes both internal AV features and various types of external information sources for event detection in team sports video. Three schemes are also proposed to tackle the asynchronism between the fusion of AV and external information. The framework is extensible as it can provide increasing functionalities given more detailed external information and domain knowledge. By demonstrating its effectiveness on soccer and American football, we believe that with the availability of appropriate domain knowledge, the framework is applicable to other team sports.

Categories and Subject Descriptors: H.2.4 [**Database Management**]: Systems—*Multimedia databases*; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video analysis*

General Terms: Algorithms, Design, Experimentation

Additional Key Words and Phrases: Sports video, semantic, event detection, event modeling

1. INTRODUCTION

Semantic analysis of video of various sports has been actively studied, including soccer [Xie et al. 2002], swimming [Bertini et al. 2003], tennis [Duan et al. 2003], and others. As a basic semantic entity, event in sports video refers to a video segment that conveys a complete and semantically interesting meaning according to game rules. Event detection is the task of identifying these video segments with regard to their individual semantic meanings. Much research has been done in event detection [Duan et al. 2003; Zhang and Chang 2002] which serves as the basis for annotation [Bertini et al. 2003], indexing, highlight generation, and summarization.

Most current research focuses on analyzing the internal audio-visual (AV) features of video to detect events. Early efforts attempted to infer events from patterns of low-level features. A more rigorous heuristic domain model-based approach [Bertini et al. 2003] extended this AV analysis approach to incorporate multimodalities of audio [Xu 2003], videotext [Zhang and Chung 2002], and automatic speech-to-text transcripts (ASR). Domain model-based approaches have been very successful in practice, as they are precise, easy to implement, and computationally efficient. However, models are laborious to construct and are seldom reusable, and they are not able to handle subtle events that do not have a distinct AV appearance such as yellow/red card events in soccer. Because of these limitations, only a subset of events in a domain can be detected using the heuristic AV-oriented domain model-based approach.

Authors' address: H. Xu and T.-S. Chua, School of Computing, The National University of Singapore, Singapore 117543; email: xuhuaxin@comp.nus.edu.sg.

As more features are incorporated, the heuristic domain model-based approach becomes more inefficient and difficult to handcraft; the machine learning approach naturally evolved to support fully-automated event detection. Zhou et al. [2000] used decision tree to model the domain knowledge of basketball in rules, while Han et al. [2002] utilized maximum entropy classifiers to find the most suitable features and models for event detection. These methods analyzed each shot independently, without taking into account their sequential relationships. Assfalg et al. [2002] recognized the role of sequential relations and modeled the event using HMMs.

To boost robustness against variation in low-level features and improve adaptability of event detection schemes, mid-level representation has also been used in event detection. For example, Duan et al. [2003] used shot classes and audio keywords as mid-level features.

In a broader interpretation, structure analysis can also be viewed as an event detection task. Though team sports videos share a structure of alternating plays and breaks, they vary in degree to which they are structured. Well controlled matches, such as those of American football and rugby league, have a clear cut structure with predictable durations of plays/breaks and consistent signaling scenes at certain times. In matches of other sports, such as soccer and hockey, players are given more freedom to compete, and the match has a looser structure. Li et al. [2001] tried to identify plays in American football using a heuristic method. To detect the play/break structure in soccer video, many methods were proposed. They basically fell into three groups. Xu et al. [2001] identified plays and breaks by scene types (long, medium, and close-up) in a heuristic way. Xie et al. [2002] and Li and Sezan [2001] modeled play and break with HMMs, classified each part of the video bounded in a sliding fixed-length window as either play or break, and concatenated adjacent parts of the same class. Though single-layer HMMs were successful in classifying a candidate segment, they could not identify the boundaries of plays or breaks before they were segmented. Xie et al. [2003] utilized HHMM to identify instances of breaks and plays and their boundaries simultaneously in an unsupervised manner.

A major trend in event detection is the intermodal collaboration of video, audio, and textual information. Sources for textual information are videotext [Zhang and Chung 2002], automatic speech recognition (ASR) transcripts, and closed caption text [Babaguchi and Nitta 2003]. Babaguchi and Nitta [2003] described a typical example of this type of system which analyzed closed captions to detect events with the help of audio and shot boundary detection.

The previous review shows that AV-based event detection systems inevitably suffer some drawbacks. Deterministic model-based approaches have problems with high dimensionality in feature space, while learning-based methods are likely to fail to discover all patterns given insufficient training samples. Effectiveness of both types of approaches is limited by discriminative power and consistency of events' AV patterns. On the other hand, Babaguchi et al. [2004] suggested utilizing gamestats, which is information external to AV streams, as cues to identify events. The system was restricted to a particular sport domain (American football) using particular external information (gamestats). As there are many sources of external information varying in sport domain, content, and format, it is worthwhile to examine them in general with respect to the potential for facilitating event detection.

Another aspect to look at in an event detection system is how much adaptation effort is required before the system can be applied to a new domain, that is, portability. Portability is desirable, as it saves the labor required to develop a different system for each domain. Nevertheless, this is difficult due to the diversity in event models across different domains. While most existing systems worked on only a single domain, a few tried to address a range of domains. Li and Sezan [2001] proposed a model template for detecting plays in several sports: American football, baseball, and wrestling. However, the model template served more as a guideline than as a concrete framework, and the features were sport-specific. Moreover, it was only capable of differentiating plays from non-plays, rather than classifying

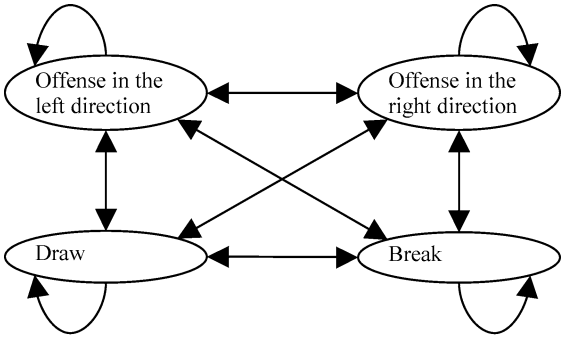

Fig. 1. Overall structure of team sports video.

among event types. Bertini et al. [2003] used modal checking to detect events in soccer and swimming. The features were common across event types and sports, but the models were event type-specific and required substantial human effort to develop.

The aim of this article is to develop an extensible framework that effectively uses both internal AV features and external information sources to detect events in broadcast team sports video. Team sports refer to the group of sports that are played by two teams moving freely on a field and scoring by manipulating a ball into fixed targets, including soccer, American football, and rugby league, etc. The main contributions in this research are a) the development of this framework and b) the fusion of multiple asynchronous intrinsic and extrinsic information sources. The framework is extensible in the sense that it can provide increasing functionalities as more external information is available. Also, the framework is portable in the sense that it may be applied to more than one team sport provided appropriate domain knowledge is available. The framework has been found to work on soccer and American football and may be able to work on more team sports.

The rest of the article is organized as follows. Section 2 explains the modeling of events in team sports video; this serves as the basis for proposing the event detection solution. Section 3 describes external information sources available for soccer and American football. Sections 4 to 7 present the proposed method. Section 8 implements the framework on soccer and American football and evaluates its performance. Section 9 concludes the article.

2. MODELING OF EVENTS IN TEAM SPORTS VIDEO

Common to all team sports videos is a distinct characteristic and that is the advancement towards two opposite targets alternately. The term *offense* refers to such advancement. Recognizing the video structure in terms of offense is beneficial to event detection as this structure suggests when certain events could happen. For example, score-related events happen mostly at the ends of the offenses (e.g., goals in soccer); launching of plays in the beginnings of offenses (e.g., punt-returns in American football), and referee interventions between offenses (e.g., yellow-card in soccer).

Modeling of events in team sports video comprises three aspects—temporal location specifications, semantic compositions, and AV patterns of event types.

Temporal location specifications of event types. Temporal locations of event types need to be studied in the context of a video's overall structure. The overall structure of a team sports video is modeled as a finite state machine with the states of *offense in particular directions*, *draw* and *break* (Figure 1). Draw describes the state when no team is on the offensive; this happens in loosely-controlled games such as soccer and hockey. Break refers to the state when the video is not showing on-going play, such

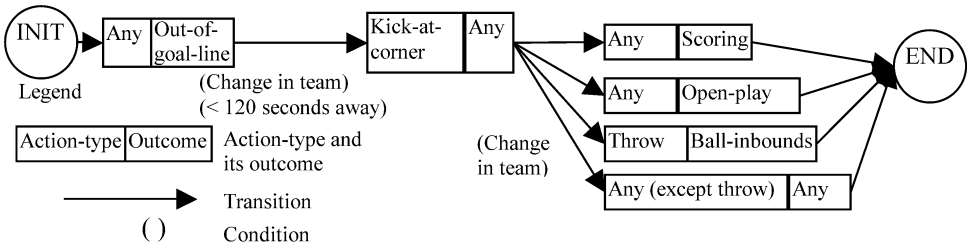

Fig. 2. Semantic composition model of corner-kick.

as when the ball is out of play or the video is showing a replay. The longest video segment that stays on a particular state is called a *phase*.

A particular event type only occurs at certain phases; this is called *temporal location specification*. For example, corner-kick in soccer only occurs in an offense that is preceded by an offense-break sequence, with the two offenses in the same direction.

Intermittent matches usually have substructures within a phase. For instance, in American football, an offense is composed of a series of plays (called downs). For these games, temporal locations may be specified on units lower than phase.

Semantic Composition Models of Event Types. Semantically, an event is composed of a series of *actions*, wherein an action refers to a single interaction between players or between players and context during the development of the event.

In terms of graph theory, a semantic composition model of an event type is an oriented graph $G{:=}(A,T)$, where A is the set of nodes representing actions with attributes, and T is the set of edges representing temporal transitions between actions. Each node $\Phi \in A$ has these attributes

$$\Phi := (\mathit{action\text{-}type}, \mathit{outcome}, \{\mathit{pre\_tran}\}, \{\mathit{post\_tran}\}); \tag{1}$$

where *action-type*[1] describes what action the player takes, *outcome* describes the state of the ball, or of related players or of the overall match as a result of the action; $\mathit{pre\_tran} := (\mathit{pre\_node},\ \mathit{pre\_cond})$ describes the transition from a preceding node into the current node and the conditions to be met during the transition; and $\mathit{post\_tran} := (\mathit{post\_node},\ \mathit{post\_cond})$ is the corresponding transition from the current into the next node; $\{\mathit{pre\_tran}\}$ and $\{\mathit{post\_tran}\}$ are the sets of pre and posttransitions.

The graph contains a number of paths connecting INIT and END. Any of these paths is a valid development of the event. For each specific event, the paths usually go through a common node which helps to distinguish the event type from others, such as the node kick-at-corner in Figure 2. Due to its distinction and consistency, it is called the key action of the event type.

AV Patterns of Event Types. Events may have patterns in audio-visual feature space. They are used as the basis to detect events in the AV space based on either a heuristic or machine-learning approach. Some event types have relatively distinct and consistent (strong) AV patterns such as goal in soccer; while others do not such as offside in soccer. Generally speaking, events' AV patterns are less consistent than semantic composition models.

[1]Some actions have linguistic terms (action types) similar to events (event types), for example, a kick-at-corner action and corner-kick event in soccer. A pair of such an action and event should be differentiated with respect to semantics and temporal range. An event describes the whole development formed by (usually) multiple adjacent actions, revealing the causality and temporal evolution; action, on the other hand, describes a single move with homogeneous semantics.

The Patriots led 20-0 at halftime, scoring on four of their five possessions, while Buffalo (3-6) punted three times and was intercepted
...
Buffalo nearly dug itself a huge hole on the first play when Terrence McGee fielded the kickoff
...
The Patriots, still rolling a 40-22 victory Nov. 7 at St. Louis, took a 6-0 lead on field goals of 27 and 24 yards by Vinatieri. On their next series, they marched 75 yards on 11 plays, capped by
...

Fig. 3. Sample of American football recap. (Source: http://www.nfl.com)

Fig. 4. Sample of soccer game log. (Source: http://www.soccernet.com)

Buffalo Bills at 15:00	
1· 10-BUF20	(15:00) W.McGahee up the middle to BUF 25 for 5 yards (T.Johnson).
2· 5-BUF25	(14:27) D.Bledsoe pass to E.Moulds to BUF 31 for 6 yards (T.Bruschi).
1· 10-BUF31	(13:47) BUF 71-Peters eligible. W.McGahee right guard to BUF 32 for 1 yard (T.Warren).
...	...
New England Patriots at 10:18	
1· 10-NE10	(10:18) C.Dillon left end ran ob at NE 13 for 3 yards.
2· 7-NE13	(9:54) C.Dillon up the middle to NE 14 for 1 yard (S.Adams).

Fig. 5. American football play-by-play report.

3. EXTERNAL INFORMATION SOURCES OF TEAM SPORTS MATCHES

External information sources are usually textual descriptions about the match, including match reports, commentaries, game logs, and so on. Since they may provide information more accurate than or supplementary to AV streams, they can be used to enhance the performance of event detection.

Despite the various sport domains and formats, we can categorize the external information sources into two levels, compact and detailed. Compact descriptions provide information about a few key events pivotal to the course of the match with coarse timing. An example of compact description is the recap in American football (Figure 3). On the other hand, detailed descriptions provide detailed information for users to follow the development of almost all events with detailed, but often slightly offset timings. Examples include the game log in soccer (Figure 4) and the play-by-play report in American football (Figure 5).

External information provides consistent semantic-rich information that AV signals lack, and helps to supplement AV analysis in event detection in two ways: (a) by providing some needed descriptors and (b) by providing the semantics directly. First, it could provide information on player's position (as shown in Figure 5), or tell the ball's position, for example, in the game log entry "08:23 Thierry Henry—Cross—Out of play". Such information cannot be reliably and accurately obtained from AV analysis

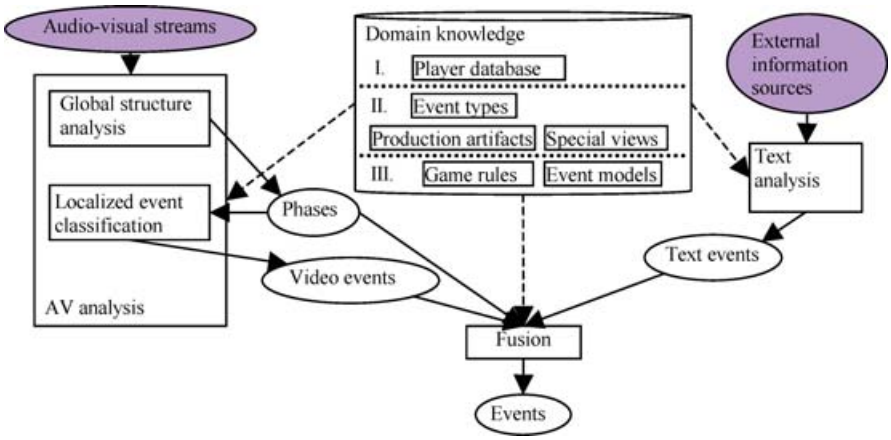

Fig. 6. The framework.

as it contains many abstract and non-AV concepts. Second, it helps by providing semantic items such as the actions shown in Figure 4. These semantic items facilitate the detection of events by semantic composition models which are more consistent than the AV patterns.

4. A FRAMEWORK FOR EVENT DETECTION

Figure 6 presents the framework of event detection in team sports by using both AV streams and external information sources. The system has three major components: (a) AV analysis which processes the AV streams, (b) text analysis which processes external information sources, and (c) fusion which combines outcomes from the two analysis. Supporting the three components is the domain knowledge which can be loaded to cater to a new team sport domain. The domain knowledge involved is mainly about game rules, player database, event models, and special visual effects. Specifically,

—*game rules* are the rules that regulate the match. Most importantly, we need to know the field layout and duration of the match;

—*player database* stores players' names, positions, nationalities, affiliations, and other background information which facilitates text analysis;

—*event models* describe the event types in the game, including choice of interesting event types, temporal units that describe events' temporal locations (phases and units lower than phase), temporal location specifications, semantic composition models, and AV patterns of event types;

—*special visual effects* refer to production artifacts, such as channel logos, used to signal replays and special views that are associated with certain content in the game like pitching in tennis video. Having special visual effects in advance facilitates AV analysis.

Though the domain knowledge involved covers a wide range of aspects, much of it can be acquired automatically. We categorize domain knowledge into three types based on the amount of manual efforts required during acquisition (see Figure 6).

Acquisition of type I domain knowledge, that is, the player database, can be fully automated. Player information can be extracted from a Web corpus using techniques similar to Yang and Chua [2004].

Type II domain knowledge is automatically acquired but needs human intervention to confirm or annotate it. Among these types are choice of event types, special views, and production artifacts. Event types are chosen by human developers from a suggested list which is extracted from a Web corpus in

the same way as player information. Special views and production artifacts are acquired by automatic clustering of video contents and subsequent manual annotation.

Domain knowledge in type III, that is, game rules and event models, is manually constructed. However, efforts in constructing event models can still be partially automated. Specially, action types and outcomes required for building events' semantic composition models are acquired in the same way as event types. Also, the AV patterns of some event types are acquired by machine-learning approaches, such as SVM.

Note that the domain knowledge in type III is the most stable part in a game, while that in type I and II is more volatile. Our data driven techniques for type I and II ensure minimization of ad hoc efforts in acquiring domain knowledge and thus enhance portability of the system.

The independence of AV and text analysis enables the system to work with various external information sources. Given different amount of external information, the system achieves different detection capabilities.

—With only intrinsic AV signals, the system achieves comparable detection capability to state-of-the-art AV-relied systems.

—By incorporating external information from compact descriptions, the system can ensure the detection of most important events, such as scorings, and those key events that do not have strong AV patterns, such as the yellow-cards and substitutions in soccer matches.

—By utilizing external information from detailed descriptions, the system can detect the full range of events and their boundaries.

5. AV ANALYSIS

From the AV streams, the AV analysis generates a list of entries call *video events*, each representing an event in terms of <start time, end time, event type>. Limited by the characteristics of AV features and the capabilities of the techniques to process them, the AV analysis only aims to detect a subset of all event types that have strong AV patterns. To achieve this aim, we adopt a two-step event detection strategy by performing (a) global structure analysis to detect phases using a statistical method and (b) localized event classification at each event-containing phase using a specific feature set and algorithm. The use of divide-and-conquer pipeline minimizes the need for more training samples and particularly alleviates the data sparseness problem. It contributes to higher precision and recall in AV event detection and ensures that our technique is able to detect a wide range of event types and deal with full-length videos.

5.1 Global Structure Analysis

To segment the video sequence into phases, the global structure analysis uses a two-layer hierarchical HMM [Xu and Chua 2004]. The top layer of HHMM models interphase transitions, and the bottom layer models the stochastic process of a phase.

To ensure the success of a learning-based approach, the judicious choice of features is important. With the aim of making the global structure analysis applicable to a wide variety of team sports, we select a set of features general to various domains, though exact definition of some features may be domain-specific. The features selected are:

—*Shot category*. It categorizes the shots into the categories of commercial, replay, audience, narrative (showing background information in videotext), and on going play.

—*Focal distance*. It classifies shots into three distinct focal distances of long-shot, medium-shot and close-up. These are found to help draw viewers' attention to important objects [Xu et al. 2002].

—*Special view category*. It tells if a shot shows a special view. Special views are defined as part of domain knowledge.

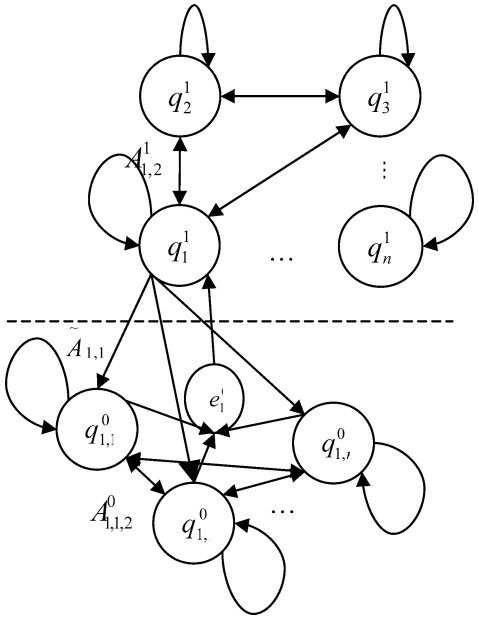

Fig. 7. Global structure analysis using HHMM.

—*Field zone*. Activities taking place at different field zones may indicate different likelihood of scoring attempts. Definitions of field zones are derived from field layout which is given in the domain knowledge.

—*Camera motion direction*. Panning and tilting help in showing the direction of movement at large and in recognizing offenses. Also, zooming implies something of great interest arises, which is usually related to scoring attempts.

—*Motion activity*. Motion activity helps in differentiating break from other phases.

We employ a range of reported techniques to detect commercials [Koh and Chua 2000], narratives [Chua and Chu 1998], replays [Chua and Chu 1998], audience [Lee et al. 2003], focal distance [Xu et al. 2000], special view category [Chua and Chu 1998], field zone [Bertini et al. 2003], camera motion direction and motion activity [Tan et al. 2000].

The general topology of the two-layer HMM is given in Figure 7. Each node q_j^1 (j=*1...n*) at the top layer denotes a phase, for example, offense in the left direction in soccer. At times, when a phase type has multiple subcategories with disparate AV patterns, each subcategory is denoted by a node. An example is return-initiated and nonreturn-initiated offenses in the left direction in American football. The number of nodes n is derived from the total number of phases or subcategories as given in the domain knowledge. The top-layer topology is ergodic since, in general, a phase can transit into any other phase or remain in the current one. Under each top layer node, there is a number of states $q_{j,k}^0$ (j=*1..n, k=1..m*) representing variations within the phase, plus an exit state e_j^0, through which the top layer node transits to another. The number of states m at the bottom layer is chosen from a number of candidate numbers that gives the best accuracy over a validation set. The top- and bottom-layer HMM are trained separately using the Baum-Welch algorithm. The Viterbi path at the top layer indicates the phases with boundaries. Global structure analysis is explained in detail in our previous work [Xu and Chua 2004].

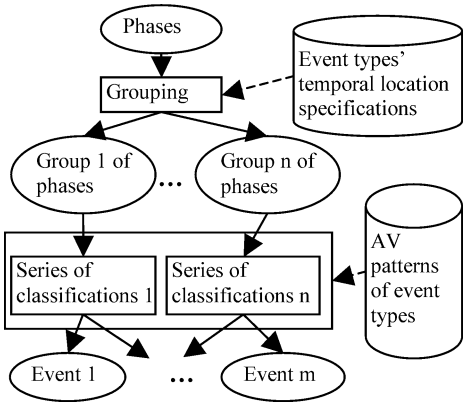

Fig. 8. Localized event classification.

5.2 Localized Event Classification

After phases are identified, those that reside in the phase sequences conforming to the same temporal location specifications are grouped. For example, a group of phases in soccer video is offenses which are sandwiched by two breaks. As a group of phases may contain more than one event type, every phase in this group are subjected to a series of classifications to determine which event type (including null) it contains. The series of classifications are based on AV patterns of the possible event types, either in a heuristic or in a learning-based method. Figure 8 illustrates this process. Details are given in Xu and Chua [2004].

6. TEXT ANALYSIS

Text analysis aims to generate a list of entries called text events, each representing an event in terms of <start time, end time, event type>. As compact and detailed descriptions have disparate contents and formats, they require different techniques to process. A description is judged as compact or detailed by a SVM classifier. The input features to the classifier are a) number of paragraphs (*PG*), b) number of time entities (*TE*), and c) number of player names (*PN*), with *TE* and *PN* normalized by the length of the article.

6.1 Processing of Compact Descriptions

Compact descriptions such as the match reports in soccer and recaps in American football are in free text form and cover only important events of interest to general readers. Due to the difficulty in processing free-form text with missing information, the list of events detected would probably be error prone and incomplete. We tackle this as an information extraction (IE) problem by using rule-based IE techniques.

Before the IE process starts, some domain knowledge needs to be put in place, namely, the player database and game rules. How to establish domain knowledge was explained in Section 4. The IE process has four steps.

(1) We induce from the training samples syntactic rules indicative of times and events, as described in Xiao et al. [2004].

(2) We identify the time entities using the rules.

(3) A window of terms $[-x, +y]$ around each time entity is picked. Usually, x and y are both set to be one sentence.

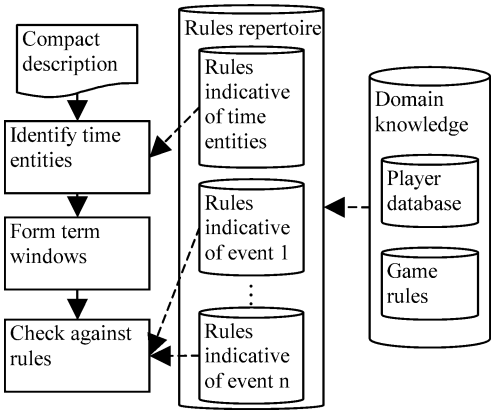

Fig. 9. Processing compact descriptions.

(4) The window of terms is analyzed against all rules to see if it satisfies any rule indicative of an event type.

When an event is found, its temporal boundaries are set to be one minute before and after the time entity, in view that the compact descriptions only give coarse timing information. Usually, one article would only mention a handful of important events; it is thus advisable to process a few (5–7) articles. In case of multiple events detected in the same term window, all are kept until the fusion module makes a decision with the help of AV analysis.

6.2 Processing of Detailed Descriptions

Detailed descriptions of team sport matches are usually composed of entries. Each entry roughly corresponds to an action, giving information on the time, action type, the player who performs it (hence possession status), and outcome of the action. Descriptions in some games may also give position information, for example, in American football, play-by-play reports. Based on information extracted from entries as well as interentry context, each entry is transformed into a node Φ:=(*action-type, outcome, pre_tran, post_tran*), which is the composing unit of an event's semantic composition model (Figure 2). It is thus logical to transform the whole detailed description to a sequence of $L = \Phi_1\Phi_2 \ldots \Phi_L$ and detect events from L by model checking. Note that entries of a detailed description could be in a field-delimited format such as game logs in soccer (Figure 4), or in free text such as play-by-play reports in American football (Figure 5). The two formats are handled differently when extracting information from entries. For the former, it is convenient to check for keywords as fields are usually filled by standard terms, whereas for the latter, rule-based IE techniques similar to those described in Section 6.1 are employed.

Having a complete sequence of $L = \Phi_1\Phi_2 \ldots \Phi_L$, occurrences of a particular event type can be identified by checking all subsequences against the model (Figure 2). A more computation-effective alternative is to perform model checking around nodes that match the key actions. From such a node, we keep crawling and analyzing preceding and following nodes as long as they satisfy the model. The last nodes leading to INIT and END, respectively, signal the detection of an event as well as mark the event's first and last actions. The event's starting boundary is taken to be the first action's recorded time, and the ending boundary is estimated to be the last action's recorded time plus the average duration of this action derived from training data.

Note that events' boundaries obtained in this way are only approximate. This is because (a) the time given in an entry is usually not accurate and (b) the duration of the last action is only an estimate. Identification of accurate boundaries requires the support of AV analysis.

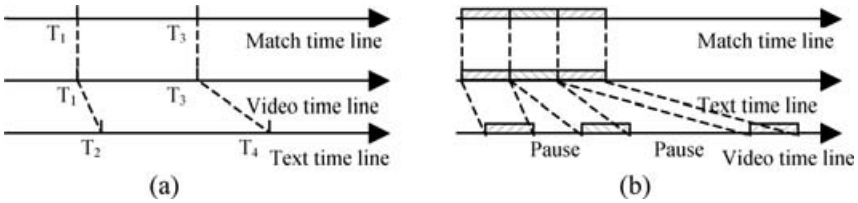

Fig. 10. Formation of offsets: (a) continuous match and (b) intermittent match.

7. FUSION OF VIDEO AND TEXT EVENTS

AV and text analysis may have conflicts in the identification of event type and boundaries because the information is presented at different semantic levels and with varying detection capabilities. The conflicts need to be resolved. Also, it would be desirable if we can take advantage of both strengths: accuracy in boundaries offered by AV analysis and reliability in detection offered by text analysis. Fusion aims to achieve these two objectives.

Many papers have been dedicated to the fusion problem for multimedia analysis, especially in a multimodal setting. Similar to the problem of sensor-fusion, fusion for multimedia analysis can be categorized into fusion of features and fusion of decisions. Fusion of features joins and transforms all relevant features before sending them as input to the classifier, such as Fisher's Linear Discriminant algorithm [Hauptman et al. 2003]. Fusion of decisions fuses outputs of individual classifiers, such as stacked SVM and bagging algorithms [Hauptman et al. 2003]. A special fusion of decisions is the fusion of rank lists. Wu et al. [2004] proposed an optimal multimodal fusion scheme by first identifying a number of independent modalities which can be viewed as fusion of features, followed by fusion of multiple modalities which can be viewed as fusion of decisions. Fusion can generally be viewed as a problem of how to derive an overall likelihood value based on a number of likelihood values. Most multimodal fusion problems studied assume that the input likelihood values refer to the same sample, that is, they are synchronized. However, this assumption may not hold in the multisource fusion problem addressed here. As we will discuss, the times associated with video and text events may not be drawn from the same clock; this fact is called asynchronism. Asynchronism obscures the temporal correspondence between a pair of video and text events. In view of this, the fusion scheme has to identify which video and text events are corresponding when determining the overall likelihood.

Here, we explain how asynchronism is formed. For text events detected from compact descriptions, asynchronism results from different time granularities of AV and text analysis. The granularity of AV analysis is frame (or millisecond), while that of text analysis is minute. Thus within the text event's boundaries, there may be multiple video events. For text events detected from detailed descriptions, asynchronism arises because a time point is recorded differently in AV and text streams. This arises in two cases for continuous and intermittent matches (see Figure 10). In continuous matches, for example, soccer and hockey, both the match and the video go on at the pace of real elapsed time, and thus they are consistent at any time point. However, the text time recorded may be different from actual video time as the human operator may need some time before he/she can tell the type and outcome of an action or he/she may anticipate a sure-fire action before it actually happens. In Figure 10(a), T_1's and T_2 refer to the same time point, and so do T_3's and T_4. The two T_1's on the match and video time lines have the same reading, which is different than T_2. The discrepancy between text time and the match time ($T_2 - T_1$) is called *offset*. The offset is usually varying ($T_2 - T_1 \neq T_4 - T_3$) and unpredictable as dictated by its underlying process.

In the case of intermittent matches, such as American football, the match pauses and resumes frequently. Given the well-bounded actions with sufficient break time when the match pauses, a human

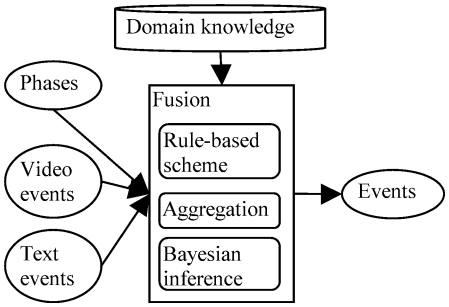

Fig. 11. The fusion process.

operator can record times of actions accurately. On the other hand, video goes on at the pace of real elapsed time with commercials, replays, and narratives inserted during match pauses (see Figure 10(b)). Therefore a time point has discrepant readings on video and text time lines, and the time line intervened with offsets in this case is the video time line. The discrepancy can be very large as the whole duration of match with pauses is 3 times the duration of the actual match (180 minutes vs. 60 minutes), which hinders the fusion process. To overcome this problem, we prepare a cleaner video as the basis for fusion by some preliminary processing—removing the commercials, replays, and narratives, linking the remaining parts and linearly scaling it to the duration of a match (60 minutes). Note that the resulting video is only an approximation of the actual match, as some scenes may appear in both pauses and plays, such as the close-ups, and cannot be removed. Also, there may be wrong or missing detections of commercials or replays, and so on. Therefore, offsets still exist.

7.1 The Fusion Process

In this research, we investigate three fusion schemes to fuse the video and text events with varying offsets (see Figure 11). They differ in the way offsets are modeled. The rule-based scheme assumes that offsets cannot be numerically modeled, whereas aggregation models offsets as following a probabilistic distribution, and Bayesian inference models them as binary.

7.2 Rule-Based Scheme

Following the guideline of "identifying the pair of items before fusing them", the rule-based scheme is accomplished in three steps.

(1) *Aligning text events and phases*. Given that there is no modeling of offsets, the alignment is sought by maximizing the number of matches between text events and phases. Here a match between a text event and a phase means that the phase conforms to the event modeling of the text event (e.g., an offense followed by a break conforms to goal's event modeling), they are both within a temporal range, and they occur in the same sequential order. As text events may overlap temporally, such as a corner-kick and a resulting goal in soccer, and multiple text events may occur in the same phase, such as a punt-return and a touchdown, respectively, in the beginning and at the end of an offense in American football, a phase may match multiple text events. The maximization problem is similar to the Longest Common Subsequence (LCS) matching problem and can be solved by the dynamic programming technique.

(2) *Determining event type*. This step resolves the conflicts in event type based on video and text events' comparative accuracy. Figure 12 gives the pseudocode.

```
for each phase aligned to a text_event
do video_event ← result of AV analysis in phase
   if video_event.type == text_event.type
      then event.type ← video_event.type
      else if F1# of AV analysis on video_event.type > F1 of text analysis on text_event.type
            then event.type ← video_event.type
            else event.type ← text_event.type
                  find video_event such that video_event.type == event.type
for each phase aligned to no text event
do video_event ← result of AV analysis in phase
   if video_event.type != null
      then if text analysis is on detailed description*
            then event.type ← null
            else event.type ← video_event.type
      else continue
for each text_event aligned to no phase
do discard text_event

# F1 = 2(precision)(recall) / (precision + recall) obtained from training data
*assume detailed descriptions give complete list of text events
```

Fig. 12. Process to determine event type.

If the fused event type is suggested only by text analysis, we proceed with additional AV analysis to find the event's location and temporal boundaries. During this process, some distinct characteristics in the events' AV patterns are utilized. Note that these characteristics may not be discriminative enough to differentiate between event types. An example of these AV characteristics is the hard-cut between an offense and a break in the AV patterns of goal, save, and shot-off-target.

(3) *Determining event boundaries*. After video and text events agree on the event type, boundaries of the fused event are determined as those of the video event.

7.3 Aggregation

As discussed earlier on, one of AV and text time lines is intervened with offset, and the other is not, for example, for soccer, the time line of text is intervened with offset and that of AV is not; for American football, the reverse is true. For simplicity, we use *offset time line/analysis/event* to respectively refer to the time line intervened with offsets, analysis conducted on the time line, and the detected events, and *nonoffset time line/analysis/event* to refer to their corresponding counterparts associated with an accurate time line.

In general, detection results of a particular event type can be depicted by a likelihood curve on a time line. The idea of aggregation is for the likelihood curve given by the offset analysis to migrate from the offset time line to the accurate time line. By doing this, the two likelihood curves given, respectively, by the offset and nonoffset analysis are synchronized and can be combined. The whole process is carried out in three steps as explained in the following.

(1) *Modeling the offset distributions for start/end of events*. This is similar to that of Yang et al. [2004] that modeled the probability of a face occurring along the time line with respect to when the name is mentioned.

(2) *Computing the likelihoods of the offset event given by the two analysis on the accurate time line*. Based on the two distributions showing probabilistically when the offset event starts or ends on the accurate time line, we calculate the probability of any time point t on the accurate time line being in the span of the offset event.

$$P_{in}(t) = \int_{-\infty}^{t} D_s(x)\,dx \cdot \int_{t}^{+\infty} D_e(x)\,dx, \tag{2}$$

where $D_s(x)$ and $D_e(x)$ are distributions of the offset event's start and end on the accurate time line, respectively.

Suppose offset event has event type i, the likelihood of event type i at time point t seen by the offset analysis on the accurate time line is:

$$P_{i-O}(t) = C_i P_{in}(t), \tag{3}$$

where C_i reflects the confidence of the offset analysis on event type i. We take it to be the precision of the offset analysis on i over the training set.

Suppose event type j is detected by the nonoffset analysis at time point t, the likelihood of event type i at time point t seen by nonoffset analysis is:

$$P_{i-N}(t) = Confusion_{ij}, \tag{4}$$

where $Confusion_{ij}$ is the element of confusion matrix that indicates the percentage of type i samples out of all samples detected to be of type j. Note that the confusion matrix includes the null event type.

(3) *Combining the likelihoods of the offset and nonoffset analysis*. Let $P_{i-N}(t)$, $P_{i-O}(t)$ and $P_i(t)$ denote the likelihoods seen by nonoffset analysis, offset analysis, and the fused likelihood on the accurate time line. Then $P_i(t)$ is computed by

$$P_i(t) = wP_{i-N}(t) + (1-w)P_{i-O}(t). \tag{5}$$

If $P_i(t)$ is greater than a threshold thr, the fused event at time point t on the accurate time line is of type i. We find the optimal parameters (w, thr) by optimizing detection accuracy over a validation set with gradient descent.

7.4 Bayesian Inference

Different from aggregation, this scheme does not model the offsets in probabilistic distribution. Instead, it only differentiates if the offset is within a maximum allowed range. Unless specified, the following description is for a particular event type p. Regarding whether p is present at time point t on the accurate time line, there is a binary hypotheses: H_0 – not-present and H_1 – present. Most likely, the hypothesis the is $\arg\max_{i\in\{0,1\}} P(x_N, x_O|H_i)\cdot P(H_i)$, where x_O and x_N are two variables derived from offset and nonoffset analysis, respectively. Usually, x_O refers to whether p is detected in the maximum allowed range, and x_N refers to the event type of detection by nonoffset analysis (it could be different than p) at time t. Since x_O and x_N are outcomes of two independent analysis, we have:

$$\arg\max_{i\in\{0,1\}} P(x_N, x_O|H_i)\cdot P(H_i) = \arg\max_{i\in\{0,1\}} P(x_N|H_i)\cdot P(x_O|H_i)\cdot P(H_i). \tag{6}$$

$P(H_i)$, $P(x_N|H_i)$ and $P(x_O|H_i)$ are obtained from the training set.

8. IMPLEMENTATION AND EVALUATION

We implement the framework on soccer and American football videos to test its portability to different domains.

8.1 Implementation on Soccer Videos

8.1.1 *Domain Knowledge.* Soccer has four phases—break, draw and offense in the left and right *directions*. The event types of interest in soccer are goal, save, shot-off-target, penalty, free-kick[2],

[2]Only free-kicks in the attacking third are considered.

Table I. Series of Classifications on Group I Phases (Soccer)

Step	Purpose	Input	Outcome	Algorithm	Features
(a)	To differentiate *goals* from *non-scoring* (union of *attempt-on-goals* and *none-of-the-group*)	*Offenses* satisfying temporal location specifications	*Goals*	By rule: No *offense* between this *offense* and subsequent presence of goal videotext.	(i) Presence of goal videotext
(b)	To differentiate *attempts-on-goal* from *none-of-the-group*	*Offenses* satisfying temporal location specifications, *goals* excluded	*Attempts-on-goals*	By rule: high excitement level in commentator's speech during the *offense* with *close-ups* following the *offense*	(i) Excitement level (ii) Presence of subsequent close-ups

corner-kick, yellow-card, red-card, substitution and offside. They are grouped according to temporal location specifications: (a) *goal/save/shot-off-target/offside*—in an offense which is followed by a break, (b) *penalty/corner-kick/free-kick*—in an offense, which is sandwiched by two breaks, and if an offense precedes the first break, the two offenses should be in the same direction, and (c) *yellow-card/red-card/substitution*—in a break in the midst of the match. In the semantic composition models of these event types, actions are modeled by a definite set of action types and outcomes. The action types are dribble, pass, cross, shoot, goal-kick, block, clear, catch, substitute[3], parry, kick-at-penalty-spot, kick-at-corner, kick-at-other-spots, throw, and unsportsmanlike-conduct. The outcomes are success-catch, failed-catch, success-block, failed-block, success-clear, failed-clear, offside[4], ball-over-bar, ball-out-of-goal-line, ball-out-of-sideline, ball-inbounds, ball-on-target, ball-hit-woodwork, scoring, players-in-attacking-third, players-in-defending-third, foul-declaration, yellow-card-issuance, red-card-issuance, open-play, and play-stop.

Soccer video has one special view—behind-goal-post—which captures activities close to the goal post from behind the goal net. Five field zones are defined based on the field layout, namely mid-field, left-corner, right-corner, left-third-except-corners, and right-third-except-corners.

8.1.2 *Domain-dependent Design and Processing.* The subset of event types detectable by AV analysis are goal, attempt-on-goal (union of save and shot-off-target), penalty, corner-kick and free-kick; while all event types are detectable by text analysis. The top layer of the HHMM for global structure analysis has four nodes, each corresponding to a phase. After experimenting with different numbers of states ranging from 2 to 9, we found that 3 states at the bottom layer give the best accuracy. Based on temporal location specifications, events detectable by AV analysis occur in two groups of phases. One is an offense which is followed by a break which may contain a goal or an attempt-on-goal; and the other is an offense which is sandwiched by two breaks which may contain a penalty, corner-kick, or free-kick. The series of classifications for the two groups of phases are shown in Tables I and II. For text analysis, the player database is built by issuing questions about teams and players to a FADA question answering engine [Yang and Chua 2004] which searches on the league's and clubs' official Web sites. Compact descriptions are processed using rule-based IE techniques, and the rules are induced by GRID technique [Xiao et al. 2004]. As detailed descriptions used for soccer, game logs, are in field-delimited format, action information is readily derived from textual templates. More details of AV analysis (including features of each classification step), text analysis, and fusion on soccer are given in Xu and Chua [2004].

[3]The substitute event is temporally and semantically equivalent to substitute action.

[4]The offside event comprises a single action with offside as the outcome.

Table II. Series of Classifications on Group II Phases (Soccer)

Step	Purpose	Input	Outcome	Algorithm	Features
(a)	To differentiate *placed-kicks* (union of *penalty*, *corner-kick*, and *free-kick*) from *none-of-the-group*	*Offenses* satisfying temporal location specifications	*Placed-kicks*	HMM	(i) Focal distance (ii) Unit duration (iii) Motion activity (iv) Camera motion direction
(b)	To differentiate among *penalty*, *corner-kick* and *free-kick*	*Placed-kicks*	*Penalties*, *corner-kicks* and *free-kicks*	Multi-class SVM	(i) Distance of the overall camera motion along the flying of the ball (ii) Angle of the overall camera motion along the flying of the ball (iii)–(v) Lengths of the three longest field lines before the flying of the ball (vi)–(viii) Angles of the three longest field lines before the flying of the ball (ix) Duration of the medium shots in the preceding break before the flying of the ball

8.2 Implementation on American Football Video

8.2.1 *Domain Knowledge.* On the match time line, American football has only two phases, offense in the left and right directions. Note that draws do not exist as there must be a team on the offensive at every moment. Breaks are removed to make a cleaner video. There is a temporal unit at a level lower than phase, play, reflecting the intermittent structure of the American football match. A play is made up of a continuous segment of match. The event types of interest are touchdown, conversion (regardless of 1 or 2 extra points and whether successful), field-goal, safety, punting, punt-return, and kickoff-return. As events are usually bound in plays, play specifies temporal locations of event types best, including: a) touchdown—in the second to the last play of an offense, which is followed by a conversion as the last play; b) conversion—in the last play of an offense, which is preceded by a touchdown as the second to the last play; c) field-goal/*safety*/*punting*—in the last play of an offense; and d) punting-return/kickoff-return—in the first play of an offense. Semantic composition models of event types are composed of action types and outcomes. Action types are tackle, pass, recover, forward-progress, backward-progress, kick, punt-kick, and scrimmage. Outcomes are ball-passed, ball-intercepted, ball-out-of-end-line, ball-out-of-bounds, ball-dropped, ball-recovered, success-tackle, failed-tackle, return, scoring-touchdown, and scoring-field-goal.

American football has two special views (a) facing-goal-post which captures a player from behind in the background of goal post, and (b) scrimmage-play which shows an on-going play started from a scrimmage scene. Field-zones in American football are mid-filed, left-end-and-red-zone, and right-end-and-red-zone.

8.2.2 *Domain-Dependent Design and Processing.* The subset of event types detectable by AV analysis are touchdown, conversion, field-goal, safety, punting, and return (union of punt return and kickoff return); while all event types are detectable by text analysis. In global structure analysis, the top layer HMM has four nodes, with two representing return-initiated and non-return-initiated offenses in each phase. Experiments with different numbers of states ranging from 2 to 9 show that 3 states at the bottom layer give the best accuracy.

In AV analysis, events are bound in plays. Play grouping is similar to phase grouping for soccer video. Two groups of plays are formed: one is the last play of an offense which may contain a conversion, a field-goal, a safety, or a punting, and the other is the first play of an offense which may contain a return. Each group undergoes a series of classifications as Tables III and IV show. Note that as touchdown and

Table III. Series of Classifications on Group I Plays (American Football)

Step	Purpose	Input	Outcome	Algorithm	Features
(a)	To differentiate among *conversion*, *field-goal*, *safety* and *non-scoring* (collective event type incorporating *punting* and *none-of-the-group*)	Last *plays* of all *offenses*	*Conversion*, *field-goal*, *safety* and *non-scoring*	By rule: particular score indicates particular event type	(i) Score update by videotext
(b)	To differentiate between *punting* and *none-of-the-group*	*Non-scorings*	*Punting* and *none-of-the-group*	SVM	(i) Distance of the overall camera motion during the *play* (ii) Angle of the overall camera motion during the *play* (iii) Presence of canonical global view

Table IV. Series of Classification on Group II Plays (American Football)

Step	Purpose	Input	Outcome	Algorithm	Features
(a)	To differentiate between *return* and *non-of-the-group*	First *plays* of all *offenses*	*return* and *non-of-the-group*	SVM	(i)–(ii) Number of *pan-lefts*/ *pan- rights* during the *play* (iii)–(iv) Distance of all *pan-lefts*/ *pan-rights*

conversion always come together, they are detected at one go, and they are identified by particular bounding plays. For text analysis, the player database for American football is built using the same method as soccer. We found it more reliable to acquire event and action lists from linguistic statistics on Web sites that provide reports or glossaries. Compact descriptions are processed using rule-based IE techniques and so are the detailed descriptions—play-by-play reports—as they are free text in each time entry.

8.3 Experimental Results

We conducted experiments to evaluate accuracy of individual modules as well as the whole system, including the global structure analysis, separate AV and text analysis, and event detection after fusion. Also, we compared performance of various fusion schemes under different conditions in the hope of finding the optimal fusion scheme.

Statistics of training and testing data for soccer and American football are summarized in Tables V and VI, respectively. All matches are full length. Match reports and game logs in soccer were obtained from www.soccernet.com, while recaps and play-by-play reports in American football were downloaded from www.nfl.com.

8.3.1 *Evaluation of Global Structure Analysis.* Since ground truths of phases' boundaries are approximate, we allow some tolerance. We denote a segmented phase by a triplet (phase category, start time t_s, end time t_e), and the tolerances of start and end time are σ_s and σ_e, respectively (both are set to be 3 seconds in our implementation). The detection is regarded as correct if its phase category is correct and the start and end times are within the tolerance range. Tables VII and VIII show that phase segmentation by 2-layer HMM performs quite well with accuracy of over 80% for soccer and over 90% for American football. We note that the segmentation results of American football are significantly better than that of soccer. This is because American football is more structured than soccer. Analysis

Table V. Statistics of Experimental Data (Soccer)

	Training	Testing
Number of matches	5	5
Total duration	460 mins	455 mins
Number of goals	13	17
Number of saves	28	43
Number of shot-off-targets	62	61
Number of penalties	2	1
Number of corner-kicks	28	37
Number of free-kicks	17	15
Number of offsides	23	28
Number of substitutions	19	15
Number of yellow-cards	8	6
Number of red-cards	1	1

Table VI. Statistics of Experimental Data (American Football)

	Training	Testing
Number of matches	5	5
Total duration on match time line	300 mins	300 mins
Number of touchdowns	30	19
Number of conversions	30	19
Number of field-goals	14	16
Number of safeties	1	0
Number of puntings	39	44
Number of punt-returns	23	20
Number of kickoff-returns	53	45

Table VII. Confusion Matrix of Phases Detected (Soccer)

	Total	(a)	(b)	(c)	(d)	(e)	Recall
L-offense (a)	403	359	3	15	13	13	0.89
R-offense (b)	352	5	310	12	10	15	0.88
Draw (c)	377	35	42	281	11	8	0.75
Break (d)	768	21	18	22	665	42	0.87
Other (e)	—	9	14	5	144	—	—
Precision	—	0.84	0.80	0.84	0.79	—	—

Table VIII. Confusion Matrix of Phases Detected (American Football)

	Total	(a)	(b)	(c)	Recall
L-offense (a)	63	60	0	3	0.95
R-offense (b)	63	0	57	6	0.90
Other (c)	—	6	7	—	—
Precision	—	0.91	0.89	—	—

shows that missing phases in soccer are mainly trivial breaks when the ball gets out of bound and short draws sandwiched by offenses in opposite directions. A significant group of false positives in soccer are goal kicks which are draws being recognized as offenses. These missing phases and false positives have limited impact on the alignment of phases and text events.

8.3.2 *Evaluation of Event Detection by Separate AV/text Analysis.* In evaluating an event detection method, criterion of correct detection of an event is defined in the same way as that of phase segmentation. Tables IX and X give the confusion matrices of events detected by AV analysis only for soccer

Table IX. Confusion Matrix of Events Detected by AV Analysis (Soccer)

	Total	(a)	(b)	(c)	(d)	(e)	(f)	Recall
Goal (a)	17	15	0	0	0	0	2	0.88
Attempt-on-goal (b)	104	0	65	0	0	0	39	0.63
Penalty (c)	1	0	0	1	0	0	0	1.0
Corner-kick (d)	37	0	0	1	25	6	5	0.68
Free-kick (e)	15	0	0	2	2	9	2	0.60
Other (f)	—	6	40	0	4	2	—	—
Precision	—	0.71	0.62	0.25	0.81	0.53	—	—

Table X. Confusion Matrix of Events Detected by AV Analysis (American Football)

	Total	(a)	(b)	(c)	(d)	(e)	(f)	Recall
Touchdown/conversion (a)	19	13	3	0	2	0	1	0.68
Field-goal (b)	16	2	10	0	3	0	1	0.63
Safety (c)	0	—	—	—	—	—	—	—
Punting (d)	44	0	0	0	32	3	9	0.73
Return (e)	65	0	0	0	6	53	6	0.82
Other (f)	—	0	0	—	5	17	—	—
Precision	—	0.87	0.80	—	0.67	0.73	—	—

Table XI. Performance of Event Detection by Text Analysis (Soccer)

	Goal	Save	Shot-off-Target	Penalty	Corner-Kick	Free-Kick	Offside	Substitution	Yellow-Card	Red-Card
Recall	0.71	0.77	0.66	1.0	0.73	0.67	0.86	0.53	0.33	0.0
Precision	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table XII. Performance of Event Detection by Text Analysis (American Football)

	Touchdown	Conversion	Field-Goal	Safety	Punting	Punt-Return	Kickoff-Return
Recall	0.53	0.53	0.69	—	0.48	0.35	0.44
Precision	1.0	1.0	1.0	—	1.0	1.0	1.0

and American football, respectively. From the tables, we observe that the precision and recall of goal in soccer and touchdown and field-goal in American football are not satisfactory. This is because our supporting videotext detection tool is not very robust, and some videotexts for scoring are mixed up with those of other types such as substitution. Some of the other event types that have poor precision or recall rates suffer from anomaly in AV features, for example, attempts-on-goal may be accompanied with controlled rather than loud cheering. Besides these observations, we already knew that AV analysis could not recognize some event types of general interest, such as red-card in soccer, and that it could not distinguish some event types with disparate semantic meanings, such as punt return vs. kickoff return in American football. These phenomena suggest that AV analysis's detection capabilities are limited by the low discriminative power or inconsistency of AV patterns.

Tables XI and XII give the precision and recall of event types detected from text analysis based on detailed descriptions for soccer and American football, respectively. For American football, we use the video that has undergone preliminary processing to test the performance of text analysis. From the tables, we observe that the precision rates of all event types in soccer and American football are perfect, whereas the recall rates are not satisfactory, especially those in American football. This is because, though text events are accurate and complete in knowing events' presence, they are generally poor in telling events' exact boundaries on the video. In soccer, virtually all missing events are those that have large offsets; in American football, the video and text time lines have fairly poor correspondence. Note

Table XIII. Event Detection Results After Fusion (Soccer)

		Rule-based	Aggregation	Bayesian inference
Goal	Recall	1.0	1.0	1.0
	Precision	1.0	1.0	1.0
Save	Recall	0.9	0.83	0.93
	Precision	0.92	0.92	0.88
Shot-off-target	Recall	0.91	0.82	0.89
	Precision	0.89	0.80	0.84
Penalty	Recall	1.0	1.0	1.0
	Precision	1.0	1.0	1.0
Corner-kick	Recall	0.89	0.65	0.89
	Precision	0.89	0.80	0.85
Free-kick	Recall	0.87	0.80	0.87
	Precision	0.93	0.86	0.87

that, if we tested text analysis using original video of American football match with no preliminary processing, the text analysis would virtually retrieve nothing correct.

8.3.3 *Comparison Among Fusion Schemes.* We compare the performance of all three fusion schemes on soccer using game log as the external information source. As aggregation and Bayesian inference only support fusion on AV-detectable event types, the comparison is restricted to a subset of all event types. Table XIII summarizes the performance of different schemes.

We observe that, compared to the rule-based scheme, the recall of aggregation on save, shot-off-target, and corner-kick are significantly lower, and the precision of aggregation on shot-off-target and corner-kick are also lower. This is probably because offsets are quite diverse and aggregation is sensitive to the diversity. To provide insights into the performance of fusion schemes, we define $\theta = R/S$ for each event type, where R is the range of offset and S is the average temporal span of the event type. θ describes how diverse the offsets are. When θ is large, there is a large overlap between the distributions describing when the event starts and ends. In this case, the probability $P_{in}(t)$ (Equation (2)) is small and keeps the combined probability $P_i(t)$ small (Equation (5)). Consequently, it becomes difficult to separate positive and negative instances.

To study the impact of θ on the performance of aggregation and Bayesian inference schemes, we conducted further experiments with varying θ. Larger diversity in offsets means higher randomness. The experiment set-up is described as follows. We manipulate the range of offsets by putting the text events at various temporal distances away from the actual occurrence, while keeping the span of events unchanged. For Bayesian inference, the maximum allowed range of offsets is kept updated. We apply this manipulation to all occurrences of all event types. In the aggregation scheme, w is kept constant, and there is an optimal threshold thr for each θ. Figure 13 depicts how the optimal threshold, precision/recall rates of aggregation, and precision/recall rates of Bayesian inference change in response to θ.

We can see from Figure 13 that as θ grows, the optimal threshold decreases, and the positive and negative instances become more difficult to separate. Consequently, the precision and recall rates decline. In contrast, the precision and recall rates of Bayesian inference almost stay constant, which means that Bayesian inference is less sensitive to a larger range of temporal offsets.

Next, we compare the performance of rule-based scheme and Bayesian inference in both soccer and American football as shown in Tables XIII and XIV, respectively. In soccer, Bayesian inference has more or less the same recall rates as the rule-based scheme, but is slightly lower in precision on some event types. This is because Bayesian inference does not differentiate the relative location of video events with regards to text event (i.e., before or after) as long as they are within the maximum offset range. Thus, multiple video events could be regarded as positive instances within the maximum range

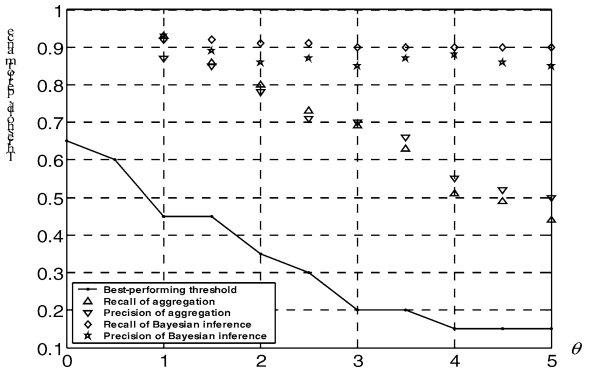

Fig. 13. Sensitivity of performance of aggergation and Bayesian inference to θ.

Table XIV. Event Detection Results After Fusion (American Football)

	Rule-Based		Bayesian Inference	
	Recall	Precision	Recall	Precision
Touchdown/conversion	0.95	0.95	1.0	1.0
Field-goal	1.0	1.0	1.0	1.0
Safety	—	—	—	—
Punting	0.94	0.93	0.96	1.0
Punt-return	0.95	1.0	1.0	1.0
Kickoff-return	0.96	0.98	0.98	1.0

from the text event, while only one of them is correct. In American football, the Bayesian inference achieved 100% for precision and almost 100% for recall over multiple event types. The outstanding performance may be explained by the fact that text events are accurate in event type and have almost zero temporal offsets. With a maximum allowable range of offset close to zero, $P(x_T \neq i|H_i)$[5] is close to zero, and thus it is almost certain that $P(x_{AV}|H_i)\cdot P(x_T = i|H_i)\cdot P(H_i) > P(x_{AV}|H_i)\cdot P(x_T \neq i|H_i)\cdot P(H_i)$, therefore $\arg\max_i P(x_{AV}|H_i)\cdot P(x_T|H_i)\cdot P(H_i) = x_T$. Since x_T is accurate, Bayesian inference also gives an accurate result. The rule-based scheme cannot achieve the same high accuracy because temporal locations are imperfect. We could see if this case still holds in soccer. Since the maximum allowed range of offset for a soccer game log is quite large, there exist some points on the video time line that are negative instances yet have $x_T = 1$. Therefore, $P(x_T = 1|H_0)$ is significant, making possible $P(x_{AV}|H_0)\cdot P(x_T = 1|H_0)\cdot P(H_0) > P(x_{AV}|H_1)\cdot P(x_T = 1|H_1)\cdot P(H_1)$ which may lead to missing some of the events.

The comparison of several fusion schemes suggests the following. Aggregation is sensitive to high diversity (or randomness) in offsets, thus it is not a reliable fusion scheme. The rule-based scheme generally performs well if the quality of alignment is good. Quality of alignment relies on two factors, accurate and complete detection of text events by text analysis which may be temporally off, and good results of phase segmentation. The rule-based scheme has the advantage of being able to detect those events that have known temporal locations but are not detectable by AV analysis. Bayesian inference has comparable accuracy to the rule-based schemes and will perform even better if the external information has offsets close to zero. However, it can only work on event types that are both detectable by AV and text analysis.

[5]For American football, x_O is x_{AV} and x_N is x_T; for soccer, x_O is x_T and x_N is x_{AV}.

Table XV. Event Detection Using Rule-Based Fusion (American Football)

	AV Only		AV + Play-by-Play Report	
	Recall	Precision	Recall	Precision
Touchdown	0.68	0.87	0.95	0.95
Conversion	0.68	0.87	0.95	0.95
Field-goal	0.63	0.80	1.0	1.0
Safety	—	—	—	—
Punting	0.73	0.67	0.94	0.93
Punt-return	0.82	0.73	0.95	1.0
Kickoff-return			0.96	0.98

Table XVI. Event Detection Using Rule-Based Fusion (Soccer)

	AV Only		AV + Match Reports		AV + Game log	
	Recall	Precision	Recall	Precision	Recall	Precision
Goal	0.88	0.71	1.0	1.0	1.0	1.0
Save	0.63	0.62	0.66	0.63	0.90	0.92
Shot-off-target					0.91	0.89
Penalty	1.0	0.25	1.0	1.0	1.0	1.0
Corner-kick	0.68	0.81	0.68	0.81	0.89	0.89
Free-kick	0.60	0.53	0.67	0.56	0.87	0.93
Offside	0	—	0	—	0.93	0.9
Substitution	0	—	0.20	1.0	0.93	1.0
Yellow-card	0	—	0.33	0.66	1.0	1.0
Red-card	0	—	1.0	1.0	1.0	1.0

8.3.4 *Evaluation of Overall Framework.* We test the performance of the overall framework on American football and soccer and use the play-by-play report, match reports and the game log as external information sources. We choose the rule-based scheme to do the fusion since we wish to detect a wide range of event types, and this is supported by reliable text analysis on detailed descriptions. The results are presented in Tables XV and XVI.

We make the following observations based on the results.

(1) Fusion of analysis of AV and compact descriptions has a trivial contribution to improving detection accuracy. This may be because the compact descriptions mention only a few events and does not help significantly in recovering events missing from AV analysis.
(2) Compact descriptions help in detecting some subtle events which cannot be detected by AV analysis, such as yellow-card/red-card, though detection of these event types is often incomplete.
(3) Compact descriptions help in ensuring detection of the most important events such as goals in soccer.
(4) Fusion of analysis of AV and detailed descriptions achieve around 90% in both recall and precision for all event types in soccer and American football. The high performance is attributed to fusion taking advantage of the strengths of both AV and text analysis. Text analysis is accurate and complete in event type identification, and AV analysis is good at boundary identification.
(5) Fusion of analysis of AV and detailed descriptions can detect the full-range of event types, including those that are undetectable by AV analysis, that is, offside, substitution, yellow-card and red-card in soccer. Moreover, fusion can differentiate events that are not differentiable by AV analysis, that is, save vs. shot-off-target, and punt-return vs. kickoff-return.

In short, the results confirm our conjecture that the framework is extensible: a) given no external information sources, the framework analyzes AV signals only and is capable of identifying the overall

structure of video in terms of phases, as well as detecting scoring-related events; b) when given compact knowledge, the framework can detect a wider range of event types and ensure detection of the most important events; and c) when given detailed descriptions, fusion can handle the full range of event types and achieves high accuracy.

9. CONCLUSIONS

We have presented a framework that fuses AV signals and external information to detect events in broadcast team sports video. We proposed three fusion schemes and compared their performance under different conditions. Experimental results showed that the rule-based scheme and Bayesian inference scheme are advisable under particular conditions, and the overall framework is effective in strengthening event detection capabilities. We also showed that the framework is extensible in that it can provide increasing functionalities when given more external information. By demonstrating its effectiveness on soccer and American football, we believe that, with the availability of appropriate sport-specific domain knowledge, the framework is applicable to other team sports.

In the future, we would like to explore the possibility of another mode of fusion which analyzes AV and textual cues in a more coherent framework and forms a single decision instead of fusing multiple decisions from separate analysis.

REFERENCES

ASSFALG, J., BERTINI, M., BIMBO, A. D., NUNZIATI, W., AND PALA, P. 2002. Soccer highlights detection and recognition using HMMs. In *Proceedings of IEEE International Conference on Multimedia and Expo* (*ICME*).

BABAGUCHI, N. AND NITTA, N. 2003. Intermodal collaboration: A strategy for semantic content analysis for broadcasted sports video. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'03)*. Barcelona, Spain (Sept.), 13–16.

BABAGUCHI, N., KAWAI, Y., OGURA, T., AND KITAHASHI, T. 2004. Personalized abstraction of broadcasted American football video by highlight selection. In *IEEE Trans. Multimedia 6*, 4, 575–586.

BERTINI, M., BIMBO, A. D., AND NUNZIATI, W. 2003. Model checking for detection of sport highlights. In *Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval* Berkeley, CA, (Nov.), 215–222.

CHUA, T.-S. AND CHU, C. 1998. Color-based pseudo-object for image retrieval with relevance feedback. In *Proceedings of the International Conference on Advanced Multimedia Content Processing*. Osaka, Japan (Nov.), 148–162.

CORMEN, T. H. 2001. *Introduction to Algorithms*, 2nd Ed. The MIT press, Cambridge, MA.

DUAN, L. Y., XU, M., CHUA, T. S., TIAN, Q., AND XU, C. S. 2003. A mid-level representation framework for semantic sports video analysis. In *Proceedings of ACM Multimedia*. Berkeley, CA (Nov.), 33–44.

GALLEY, M., MCKEOWN, K., FOSLER-LUSSIER, E., AND JING, H. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of 41st Annual Meeting of the Association for Computational Linguistics (ACL'03)*. Sapporo, Japan (July).

HAN, M., HUA, W., XU, W., AND GONG, Y. 2002. An integrated baseball digest system using maximum entropy method. In *Proceedings of ACM Multimedia*.

HAUPTMAN, A., BARON, R. V., CHEN, M.-Y., CHRISTEL, M., DUYGULU, P., HUANG, C., JIN, R., LIN, W.-.H., NG, T., MORAVEJI, N., PAPERNICK, N., SNOEK, C. G. M., TZANETAKIS, G., YANG, J., YAN, R., AND WACTLAR, H. D. 2003. Informedia at TRECVID 2003: Analyzing and searching broadcast news video. *TRECVID'03*. Gaithersburg, MD, (Nov.).

KOH, C.-K. AND CHUA, T.-S. 2000. Detection and segmentation of commercials in news video. Tech. rep. The School of Computing, National University of Singapore.

LEE, M. H., NEPAL, S., AND SRINIVASAN, U. 2003. Edge-based semantic classification of sports video sequences. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'03)*. Baltimore, MD, (July).

LI, B. AND SEZAN, M. I. 2001. Event detection and summarization in sports video. In *proceedings of IEEE Workshop on Content-based Access of Image and Video Libraries.* Kauai, HA (Dec.), 132–138.

PAN, H., LI, B., AND SEZAN, M. I. 2002. Automatic detection of replay segments in broadcast sports programs by detection of logos in scene transitions. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'02)*. Orlando, FL (May).

TAN, Y.-P., SAUR, D. D., KULKARNI, S. R., AND RAMADGE, P. J. 2000. Rapid estimation of camera motion from compressed video with application to video annotation. *IEEE Trans. Circuits Syst. Video Techn. 10*, 1, (Feb.).

WU, Y., CHANG, E., CHANG, K.C.-C., AND SMITH, J. R. 2004. Optimal multimodal fusion for multimedia data analysis. *ACM Multimedia*, (Oct.).

XIAO, J., CHUA, T.-S., AND LIU, J.-M. 2004. Global rule induction for information extraction. *Int. J. Artificial Intell. Tools 13*, 4, 813–828.

XIE, L., CHANG, S. F., DIVAKARAN, A., AND SUN, H. 2002. Structure analysis of soccer video with hidden Markov models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'02)*. Orlando, FL (May).

XIE, L., CHANG, S. F., DIVAKARAN, A., AND SUN, H. 2003. Unsupervised discovery of multilevel statistical video structures using hierarchical hidden markov models. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'03)*. Baltimore, MD (July).

XU, H. AND CHUA, T.-S. 2004. The fusion of audio-visual features and external knowledge for event detection in team sports video. In *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR'04)*. New York, NY (Oct.).

XU, H., FONG, T.-H., AND CHUA, T.-S. 2005. Fusion of multiple asynchronous information sources for event detection in soccer video. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'05)*. Amsterdam, The Netherlands, (July).

XU, M. 2003. Content-based sports video analysis using multiple modalities. MSc Thesis, National University of Singapore.

XU, P., XIE, L., CHANG, S. F., DIVAKARAN, A., VETRO, A., AND SUN, H. 2001. Algorithms and system for segmentation and structure analysis in soccer video. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'01)* Tokyo, Japan, (Aug.).

YANG, H. AND CHUA, T.-S. 2004. FADA: Find all distinct answers. In *Proceedings of WWW'04*. 304–305.

YANG, J., CHEN, M.-Y., AND HAUPTMANN, A. 2004. Finding person X: Correlating names with visual appearances. In *Proceedings of the 3rd International Conference on Image and Video Retrieval*. Dublin, Ireland (July).

ZHANG, D. AND CHANG, S. F. 2002. Event detection in baseball video using superimposed caption recognition. In *Proceedings of ACM Multimedia*. Juan-les-Pins, France (Dec.). 315–318.

ZHANG, H., LOW, C. Y., AND SMOLIAR, S. W. 1995. Video parsing and browsing using compressed data. *Multimedia Tools Applica. 1*, 89–111.

ZHANG, Y. AND CHUA, T.-S. 2000. Detection of text captions in compressed domain video. I*n Proceedings of the ACM Workshop on Multimedia Information Retrieval*. CA (Nov.), 201–204.

ZHOU, W., VELLAIKAL, A., AND KUO, C. C. J. 2000. Rule-based video classification system for basketball video indexing. In *Proceedings of the ACM Multimedia Workshops*. Los Angeles, CA, (Oct.), 213–216.

Received November 2005; accepted November 2005