Sports Event Detection using Temporal Patterns Mining and Web-casting Text

Minh-Son Dao and Noburu Babaguchi
Media Integrated Communication Lab
Graduate School of Engineering
Osaka University
2-1 Yamadaoka, Suita, Osaka 565-0871, Japan
{dao, babaguchi}@nanase.comm.eng.osaka-u.ac.jp

ABSTRACT

Event detection is one of the essential tasks by which the performance of sports video content analysis and access becomes more efficient and effective. Among internal informations which are extracted from inside raw videos, the temporal information is critical to convey event meaning. In this paper, the new method for adaptively detecting event based on Allen temporal algebra and external information support is presented. The temporal information is captured by presenting events as the temporal sequences using a lexicon of non-ambiguous temporal patterns. These sequences are then exploited to mine undiscovered sequences with external text information supports by using class associate rules mining technique. By modeling each pattern with *linguistic part* and *perceptual part* those work independently and connect together via *transformer*, it is easy to deploy this method to any new domain (e.g baseball, basketball, tennis, etc.) with a few changes in *perceptual part* and *transformer*. Thus the proposed method not only can work well in unwell structured environments but also can be able to adapt itself to new domains without the need (or with a few modification) for external re-programming, re-configuring and re-adjusting. Experimental results that are carried on more than 30 hours of soccer video corpus captured at different broadcasters and conditions as well as compared with well-known related methods, demonstrated the efficiency, effectiveness, and robustness of the proposed method in both offline and online processes.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Abstracting methods, Indexing methods*

General Terms

Algorithms, Measurement, Performance, Experimentation

AREA'08, October 31, 2008, Vancouver, British Columbia, Canada.

Keywords

Event Detection, Data-mining, Temporal Sequential Patterns, Web-casting text

1. INTRODUCTION

The timely delivery of news, information, and entertainment is a crucial factor in the development of the modern e-society. Within decades, a new revolution is going to happen in this field, due to the always increasing use of digital technologies and data networks in the production and distribution of contents. In particular, digital terrestrial television (and satellite television earlier) introduced a new concept of the broadcast services, where the user not only can personalizedly access information but also interact with the system instead of being a passive subject.

The sports broadcasting area is concerned as a commercial area which has been developing tremendously due to the always increasing use of digital technologies and data networks in the production and distribution of contents as well as audiences' requirements of entertainment. In particular, the creation of a document to be broadcasted requires shorter processing times (and then, higher automation), higher attractiveness (to cope with the increased offer), the compliance with a number of advanced requirements (such as interactivity, personalization, etc.), the scalability (to address different users, with different needs and cultures, and accessing the service through heterogeneous terminals and connections). Therefore, in order to satisfy these requirements, new advanced tools for accessing updated information, searching and composing multi-media data, creating scalable documents easily and interactively at diffirent levels of depth are expected urgently.

In fact, a major difficult in implementing such tools is that the sources (databases) from which the user can extract the requested data are usually overcrowded, geographically distributed, not properly updated, poorly indexed, and often stored in unsuitable format (either for the physical support or for the data archiving format). This makes difficult oftenly to access useful information, and to reuse available data, thus increasing time and cost for new productions.

Due to the semantic gap between low-level and high-level semantic features, many open problem related to event detection still exists. To deal with the gap, most of the existing methods are using supports from the domain knowledge. However, those methods relies heavily on the domain knowledge with signicant human interference, it can be hardly applied as a generic framework for an arbitrary domain au-

tomatically [15, 11, 2, 19]. Beside that, although temporal information is critical to convey event meaning, a few studies have been attempted to exploit it in event detection. Allen[1] proposed the temporal algebra by which any temporal relation in reality can be modelled optimally based on thirteen basic relationships, however, very little attention has been paid to utilized Allen's theory so far. Most of methods dealing with capturing temporal information are mainly on only linear temporal relations, although, they lack the ability to represent temporal information of real complex events compared with Allen's. The following part is a compact review for related methods which the temporal information is taken into account.

In [9] and [20] the authors capture the temporal information between a certain pair of states as a simple sequential relation (e.g. followed by) by building the Finte State Machine. Although using simple sequential patterns, a high-level concept that can not be detected and recognized from using solely visual descriptors without temporal relation information could be detected with high accuracy.

In order to offer strong generality and extensibility with the capability of exploring representative event patterns with little human interference, Chen et al [4, 3] propose a method using an advanced temporal analysis, and multi-modal data mining method. The basic idea of this method is that a signicant temporal pattern has ability of separating the event units as far away from the nonevent units as possible, while grouping the consecutive event units themselves as closely to each other as possible. Clearly, due to using sequential pattern this method has no ability to detect complex events in the case temporal patterns displayed at different sequential orders.

Zhu et al [22] propose the attractive method by which most of aforementioned problems are solved. The authors use multilevel sequential association mining to explore associations among the audio and visual cues, classify the associations by assigning each of them with a class label, and use their appearances in the video to construct video indexes. Unfortunately, since these tags are orderly sorted as "followed by", it is easy to see that the knowledge of temporal intervals is not maintained but simplified.

Snoek et al [14] utilize the Allen temporal algebra detect event in two different domain: sports and news. The triplex *(TIME segmentation, TIME relations, TIME representation)* is used to capture and model the temporal intervals. Unfortunately, by predefining the semantic rules for a set of limited events, this method is still inflexible and domain dependent.

These aforementioned methods have the common disadvantage: due to totally depending on the knowledge of one specific domain and human interference, these models can not be deployed to detect events at other domains, or even other events in the same domain.

In [7, 8], the authors introduce the very interesting unsupervised method detect events based on Allen's theory. Under this frame work, complex events are modelled using a lexicon of hierarchical patterns of movements whose temporal structures are represented as they are in reality, which are mined from a large corpus of unannotated video data. The result of this method is then integrated text-modal (e.g. closed caption) to build the Grounded Language Model (GLM) which maps query terms to the non linguistic context to which they references, then the system to index automatically a large corpus of broadcast baseball games using an unsupervised content-based approach is constructed. The success of this frame work is by focusing on representation the temporal structure of complex events, and taking the benefit of external-information (e.g. closed caption) as the cue to create predefined event types automatically.

Nonetheless, the study of Wu et al [18] did point out two ambiguous problems of Allen temporal algebra: (1) the same relationship among events can be represented by different temporal pattern; (2) the same temporal pattern can represent different relationships among events. This affects the accuracy of results of methods those deploy Allen' temporal algebra to represent the complex events, especially when these events are combined from several attributes streams. In order to clarify these ambiguous, Wu et al [18] proposed the new temporal pattern presentation to overcome these problems by recording the duration of each event by using its start and end time-stamp instead of treating it as one instance (see Figure 1).

In light of these discussions, this research proposes the new method for adaptively detecting event based on modified Allen temporal algebra and external information support is presented. The temporal information is captured by presenting events as the temporal sequences (TSs) using a lexicon of non-ambiguous temporal patterns (NATP) [18]. These sequences are then used to mine undiscovered sequences with external-text information supports by using technique of mining class association rules(CAR) [21].

The main contributions of this work are:

1. To the best of our knowledge, this is the first time Allen-based non-ambiguous temporal patterns mining and web-casting text support are integrated to detect sports events. Moreover, NATP are studied and modified to mine class association rules instead of association rules.

2. Thanks to the support of web-casting text, events in both training and result database are annotated automatically.

3. Most of undiscovered temporal patterns is discovered and annotated not only in training stage but also in running stage. In other words, our method has the ability to update its knowledge throughout its activities.

4. By modeling each pattern with *linguistic part* and *perceptual part* those work independently and connect together via *transformer*, it is easy to deploy this method to any new domain with a few changes in *perceptual part* and *transformer*.

5. The proposed method not only can work well in unwell structured environments but also can able to adapt itself to new domains without the need (or with a few modification) for external re-programming, re-configuring and re-adjusting.

2. PROPOSED METHOD

The proposed method is constructed as two main independent tasks: (1) temporal pattern mining; and (2) video and text analysis. These two independent tasks are linked together via so-called *transformer* that will be explained later,

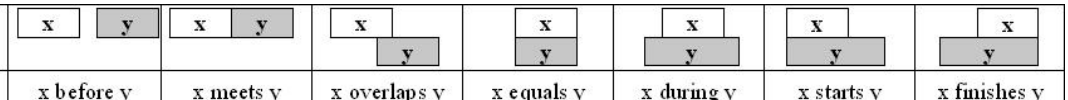
Pictorial example							
ATR	x before y	x meets y	x overlaps y	x equals y	x during y	x starts y	x finishes y
WTR	$x^{+1}<x^{-1}<y^{+1}<y^{-1}$	$x^{+1}<x^{-1}=y^{+1}<y^{-1}$	$x^{+1}<y^{+1}<x^{-1}<y^{-1}$	$x^{+1}=y^{+1}<x^{-1}=y^{-1}$	$y^{+1}<x^{+1}<x^{-1}<y^{-1}$	$x^{+1}=y^{+1}<x^{-1}<y^{-1}$	$y^{+1}<x^{+1}<x^{-1}=y^{-1}$

Figure 1: Temporal relations (ATR: Allen's temporal relation, WTR: Wu's temporal relation)

to detect events in from unknown raw video both in offline case where the video is from storage equipments such as HDD, Videotape, DVD, etc., and online case where video as an online streaming directly from broadcasters.

The mission of the former is to capture and represent the temporal intervals among basic events as temporal sequences that enable to maintain the knowledge of temporal intervals of complex target events. In this task, only the conceptual events (i.e. *linguistic*) are concerned. Therefore, this task is independent domain-knowledge. In other words, only the name of events (both basic and target) and the time intervals where those events happen are concerned (See Figure 2).

The purpose of the latter is to extract features and build visual events (i.e. *perceptual*) from a certain domain. In our study, soccer domain is chosen as the case study due to the loose structure of video, the diversification of events, and the high random occurrence of events.

3. TEMPORAL PATTERN ANALYSIS AND MINING

Assuming that there is a database of given events whose contents are the set of patterns that occur in various time periods. Let D denote that database and

$$I = (target\_event\_id, event\_id, event\_interval)$$

denote D's item where

$$event\_id = (event\_type, event\_property)$$

be a basic event and

$$event\_interval = (time\_start, time\_end)$$

be the time interval where such basic event happens. The *event_type* is presented as symbolic, called *linguistic part* and the *event_property* captures the event's *perceptual part* that model, in this case, is the visual patterns of the specializations of the concepts of the linguistic part. In other words, each real complex event (*target_event_id*) is recored as the set of basic events (*event_id*) that occur in complex and varied temporal relations to each other. These basic events can be low-level features, mid-level features or even high-level concepts those can be easily extracted from draw video clip using multimedia processing techniques or inferred from simple semantic cues .

The mission of the proposed method is to answer the following questions:

1. how to capture and represent the temporal intervals among these basic events as temporal sequences that enable to maintain the knowledge of temporal intervals of complex events.

2. how to discover temporal sequences that occur frequently in temporal database D.

3.1 Temporal patterns analysis

As mentioned in previous sections, the study of Wu et al [18] did point out two ambiguous problems of Allen temporal algebra [1]: (1) the same relationship among events can be represented by different temporal pattern; (2) the same temporal pattern can represent different relationships among events. Moreover, Allen's temporal relations identify the relationships between two intervals rather than those among more than two intervals. Therefore, if there are n intervals ($n > 2$), then the temporal relationship can not be precisely described as the sequential lexicon of (n-1) Allen's thirteen relations, but as the multi-level nested relations form, then the constructing temporal sequences becomes the burden of complexity, [7] is the best example for this situation.

In order to resolve the problem of aforementioned problems, Wu et al [18] proposed the new temporal pattern presentation to overcome these problems by recording the duration of each event by using its start and end time-stamp instead of treating it as one instance (see Figure 1). Due to the lack of space, only important definitions and algorithms those characterized the non-ambiguous temporal pattens and those are developed in the proposed method are presented, others please refers to [18] for details.

Definition 1 (Event endpoints and order relation). An event e_i has two end point e_i^+ and e_i^-, called event end points, where e_i^+ is the starting point *(esp)* and e_i^- is the ending point *(eep)* of e_i. Let the time of an event end point u, either *esp* or *eep*, be denoted as *time(u)*. Then, the order relation *Rel(u, v)* of two event end points u and v can be defined as *"<"* if *time(u) < time(v)* and as *"="* if *time(u) = time(v)*.

Definition 2 (Arrangement of event end points in a temporal sequence). In a temporal sequence (defined in definition 3), end point u must be placed before end point v if the following conditions are satisfied:

1. *time(u) < time(v)*
2. *time(u) = time(v)*, but u is an *esp* and v is an *eep*
3. conditions 1 and 2 are tied, but u's event type alphabetically precedes that of v
4. conditions 1, 2 and 3 are tied, but the occurrence number of u is smaller than that of v.

Definition 3 (Temporal sequence). Suppose there is a total of u event types that could occur in D, called *event types* 1, 2, ..., t. A temporal sequence can be constructively defined as follows:

1. A temporal sequence of one event, called a 1-event temporal sequence, can be written as $\left(e_i^+ \oplus e_i^-\right)$, where $e_i \in \{1, 2, ..., t\}$, and $\oplus \in \{<\}$ is the order relation of e_i^+ and e_i^-

2. Let $p = (p_1 \oplus_1 \cdots p_i \oplus_i p_{i+1} \cdots p_j \oplus_j p_{j+1} \cdots \oplus_{2k-1} p_{2k})$ denote a temporal sequence of k events (with $2k$ event endpoints), called a k-event temporal sequence, where p_i is an end point, and $\oplus_i \in \{<, =\}$ for all i. Suppose $p^{'}$ is the temporal sequence obtained by inserting event e_a into p. Assume that e_a^+ must be placed between p_i and p_{i+1} and e_a^- must be placed between p_j and p_{j+1} according to the rules given in definition 2. Then, we have $p^{'} = (p_1 \oplus_1 \cdots p_i \oplus_{i'} e_a^+ \oplus_{a+} p_{i+1} \cdots p_j \oplus_{j'} e_a^- \oplus_{a-} p_{j+1} \cdots \oplus_{2k-1} p_{2k}$
where $Rel(p_i, e_a^+) = \oplus_{i'}$, $Rel(p_{i+1}, e_a^+) = \oplus_{a+}$, $Rel(p_j, e_a^-) = \oplus_{j'}$, and $Rel(p_{j+1}, e_a^-) = \oplus_{a-}$

Definition 4 (Small operation). Function

$$\text{Small}(\oplus_r, \oplus_{r+1}, \cdots . \oplus_q),$$

where $\oplus_i \in \{<, =\}$, will output "<" if any $\oplus_i$, $r \le i \le q$, is "<". Otherwise, the output of Small is "=".

Definition 1, 2, 3 and 4 are the characteristics of nonambiguous temporal patterns representation. Based on these definitions, the data will be transformed from the attribute temporal database D to temporal sequence form completely, meanwhile the knowledge of temporal intervals is maintained perfectly. Thus, the *transformer* function (mentioned in section 2) of the proposed method is constructed based on these definitions.

Definition 5 (Containment). Suppose we have two temporal sequences $ts = (s_1 \oplus_1 s_2 \cdots \oplus_{n-1} s_n)$ and $p = (p_1 \otimes_1 p_2 \cdots \otimes_{r-1} p_r)$, $r \le n$. Temporal sequence p is contained in temporal sequence ts, denoted as $p \subseteq ts$, if the following conditions are satisfied:

1. There are r indices in ts, denoted as $1 \le w_1 < w_2 < \cdots < w_r \le n$, satisfying the condition of $p_1 = s_{w_1}, p_2 = s_{w_2}, \cdots, p_r = s_{w_r}$.
2. $\otimes_i = Small(\oplus w_i, \cdots, \oplus_{w_{i+1}-1})$ for i=1 to r-1.
3. For every event e in ts, the two end points of e are either both in p or not in p at all.

Definition 5 is used to determine whether whole or part of a certain temporal sequence is contained by another, and vice versa. This is useful for defining the *similarity measure* by which the similar degree between two events will be evaluated. We inherit this definition to build the *similarity measure function* for our method (defined later).

3.2 Temporal patterns mining

In [18], the TprefixSpan is designed based on well-known PrefixSpan algorithm [12] to mining temporal sequences in temporal database D. The PrefixSpan is known as a more efficient algorithm for mining sequential patterns comparing with Generalize Sequential Pattern (GSP) and Apriori. Pei et al [12] introduces the term of *Prefix* and *Projection* by which the most time-consuming task that generates and counts candidate sequences in GSP and Apriori algorithms, is avoided. Since using projection, the database PrefixSpan scans every time is much smaller than the original database, thus PrefixSpan can handle the quite long sequential pattern more efficiently than GSP and AprioriAll. Since the original PrefixSpan only deals with point-based events (i.e. each event is represented by one end point), it can not be deployed directly to interval-based events (i.e. each event is represented by two end points). Therefore, Wu et al [18] proposes the PrefixSpan-based algorithm, called TPrefixSpan (TprefixSpan algorithm could be referred in [18], for detail).

4. VIDEO AND TEXT ANALYSIS

As mentioned before, each basic event in attributes temporal database D has two parts: linguistic part and perceptual part. Meanwhile only the linguistic part is a conceptual object that is represented by a certain symbolic (i.e. acronym of event's name), the perceptual part represents the visual/audio/textual patterns that merely duplicates of things observed in reality. Thus, the mission of this section is to describe (1) how to extract required visual patterns to construct perceptual parts; and (2) how to link the perceptual part to the linguistic part.

In our study, soccer domain is chosen as the case study due to the loose structure of video, the diversification of events, and the high random occurrence of events.

4.1 Video Analysis

4.1.1 Camera motion classification

In sports, the camera motion also plays an important role. The spirit of sport broadcasting programs depends totally on the idea of director (e.g., arrangement of camera positions, number of camera, etc) and the action of camera men (e.g. following players, referee, or focusing in audience, etc). Therefore, the camera motion itself contains a copious knowledge related to the action of whole match. In the light of these discussions, in this proposal, the camera motion is treated like one of basic events combined to complex sports event. Camera motion is detected and classified by using algorithm proposed in [22]. In our case, only four types of camera motion are chosen: *Pan*(left and right), *Zoom*(in and out), *Title*(up and down), and *Unknown* (camera motions those are not Pan, Zoom, or Tilt are grouped to Unknown).

4.1.2 View classification

In [13], authors pointed out that field sport broadcasts all share common characteristics such as for each match of any sport broadcast there are three well-defined styles of camera shot: global (main), zoom-in, and extreme close-up. The above comment was explained in detail as follow: (1) *Long Shot group* (global): A long shot displays the global view of the field; (2) *In-Field Medium Shot group* (zoom-in): A medium shot, where a whole human body is usually visible, is a zoomed-in view of a specific part of the field; (3) *Close-Up Shot group* (extreme close-up): A close-up shot shows the above-waist view of one person; and (4) *Out of Field Shot group* (extreme close-up): The audience, coach, and other shots are denoted as out of field shots. From those studies, we classify the field view type into four classes: *Long view*, *Medium View*, *Close-up*, and *Out-of-Field*. In order to classify those view type, first we use the algorithm of Liu et al [10] to detect play field zone. Then the method in [6] is used to classify view types.

4.1.3 Arc and Goal Position extraction

Goal position, middle field circle and penalty box arc play the most important roles in soccer game. They could be

semantic cues by which the complex event could be determined. For example, if the left penalty box arc and goal mouth appear in current frame parallel with fast zoom-in and pan left, it could be the cue for an attack. We use the method of Wang et al [17] to detect arcs and of Duan et al [5] to detect goal position.

4.1.4 Game Time Recognition

In most of sports video, a video clock is used to indicate the game lapsed time. This is the most important visual cue by which the alignment between the text information offered from external-text information (e.g. ASR, closed caption, or web-casting text) and the visual information are performed more accuracy. In our method, the method [20] that uses Temporal Neighboring Pattern Similarity (TNPS) measure and template sample technique to detect, is inherited to recognized the game time.

4.1.5 Replay detection

Replay is also the good semantic cue to detect events due to frequently happening after every event in sports game, especially in soccer domain. Therefore, we use *Replay* as one basic event in our method. The replay shot is detected by taking into account the fact that before and after replay shot there usually is very short shot of a flying logo of a certain broadcaster or an organization. The method in [16] is utilized to detect replay in our method.

4.2 Text Analysis

There are two external text sources that can be used for sports video analysis: closed caption, and web-casting text. The former is a transcript from speech to text and only available for certain sports games and in certain countries. The latter whose content is focused more on events of sports games with a well-defined structure and is available in many sports websites. The content of web-casting text could be classified into two groups: (1) well defined syntax structure syntax (e.g. www.uefa.com); and (2) freestyle text (e.g. www.soccernet.espn.go.com). Clearly, the former is the wise selection for extracting information due to easy extracting keyword precisely with simple keyword based text search techniques. Since the web-casting text is usually available online and provided freely by almost all broadcasters, we decide to choose web-casting text as the external-information supports.

Keywords by which events are labeled are first predefined, then the time stamp where event happens are extracted from well defined syntax structure syntax web-casting texts by using the keywords as input query key to commercial software, namely **dtSearch**[1]. This software is the strong text search engine that has ability to look-up word feature with detailing the effect of fuzzy, phonic, wildcard, stemming and thesaurus search options.

5. EVENTS CLASSIFICATION

5.1 Transformer

The following example shows how the *transformer* transform data from the attribute temporal database D to non-ambiguous temporal sequences. The table 1 denotes that at the time interval (10,100) one target event *Corner* happened. This target event has basic events *Pan left*, *Pan right*, and *Zoomin* happenning in varied time. The *transformer* uses definition 1, 2, 3 and 4 and the *codebook* that is denoted in table 2 to construct the non-ambiguous temporal sequences. Thus, the result of transformer function with input is the set of basic event and their intervals of one target event is $A = (a^{+1} < b^{+1} = c^{+1} < a^{+2} = b^{-1} = c^{-1} < a^{-1} < a^{-2}$. It should be note that basic event *Pan left* happens twice at different time intervals. Therefore, the number follow "+" or "-" is the *occurrence number* that denotes the frequency of event's occurrence. The above lexicon of symbolic is the non-ambiguous temporal sequence of target event A.

Table 1: Attribute Temproral Database D

target_event_id	event_id	event_interval	
		time_start	time_end
Corner	Pan left	10	80
Corner	Pan right	30	60
Corner	Zoom in	30	60
Corner	Pan left	60	100
...	...	.	...

5.2 NATP Mining

This subsection describes how the method works to construct the training database by which the temporal events are captured, mined, and classified into suitable groups. Unlike other methods where a training database is created and annotated manually or semi-automatically, the proposed method, with the supports of web-casting text information, such a training database is constructed automatically. The proposed method's process is described as follows: (See Figure 2 for visualization)

1. Key-words by which events are labelled are extracted from web-casting text by using the commercial software, namely dtSearch. Then, necessary features are extracted from raw videos and mapped to temporal pattern form as described in Table 2.

2. By linking the time stamp in the text event that extracted from Web-casting text to the game time in the video detected by clock digits recognition technique [20], the moment where the event happens is detected. Then, the event boundary is extracted loosely by slide the time backward and forward from the timestamp in time interval *t*. These events' contents are then decomposed into set of patterns using suitable methods and *transformer* and the *codebook* denoted in Table 2. At the end of this step, under NATP scheme, each event α and its label y is treated as ruleitem $r(\alpha,y) = (\alpha_1 \oplus_1 \cdots \oplus_{n-1} \alpha_n < y^+ < y^-)$ where $(y^+ < y^-)$ is a conceptual representation of label (i.e. it does not convey any temporal information). Results from this step are used to constructed the training temporal patterns database (TTPD).

3. The modified NATP algorithm is used to mine all temporal patterns from TTPD. Since each event is treated as *ruleitem*, mining class association rules [21] is used instead of mining normal association rules to mine temporal patterns. The major difference between

[1]www.dtSearch.com

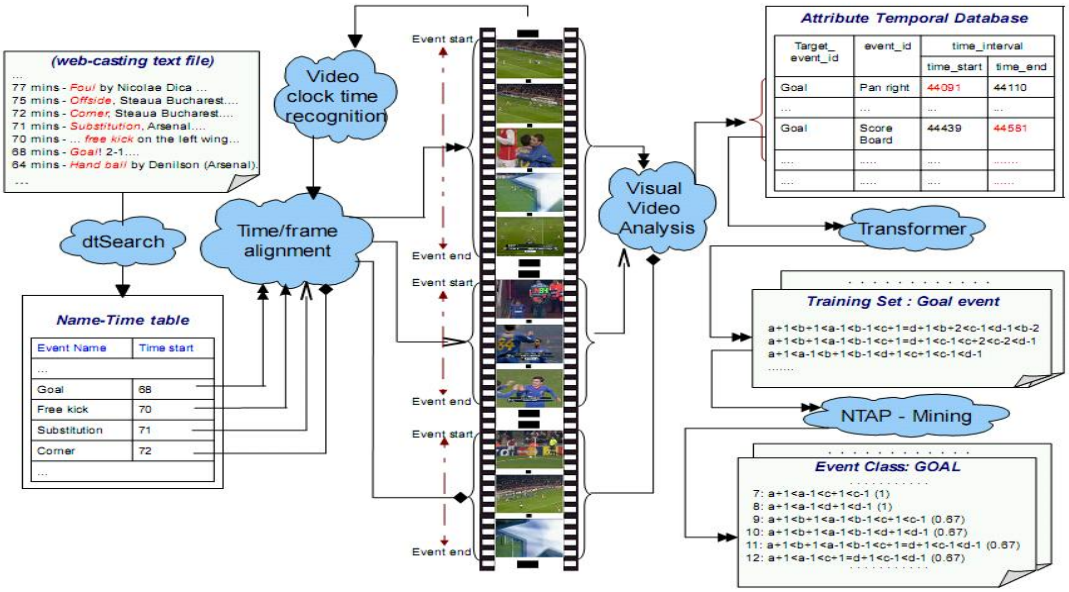


Figure 2: Event classification workflow

Table 2: Features and their NTAP

Features		Temporal Pattern	Features	Temporal Pattern
Camera motion	Pan left	$a^+ < a^-$	Pan right	$b^+ < b^-$
	Zoom in	$c^+ < c^-$	Zoom out	$d^+ < d^-$
	Tilt up	$e^+ < e^-$	Tilt down	$f^+ < f^-$
	Still	$g^+ < g^-$	Unknown	$h^+ < h^-$
View	Long view	$i^+ < i^-$	Medium View	$j^+ < j^-$
	Close up	$k^+ < k^-$	Out of field	$l^+ < l^-$
Others	Arc	$m^+ < m^-$	Replay	$n^+ < n^-$
	Goal mouth	$o^+ < o^-$	Middle circle	$p^+ < p^-$

these two methods is the former records only frequents that are mined from the training database those have class label at the end of their formulation.

6. EVENT DETECTION

We now turn to the problem of event detection from an unknown raw video both in offline case (e.g. video from storage equipments such as HDD, VideoTape, DVD, etc) , and online case (e.g. video as an online streaming from the Internet).

Let $S = S_1 \cup S_2 \cdots \cup S_n$ denote the classified temporal sequences database resulted from previous mining step, where S_i is a subset of S and contains TSs those have the same label i, $1 \leq i \leq n$. Let *maxlength* describe the number of patterns of the longest TS in S (in our method, the length of TS is counted by number of its patterns), and U represent an unknown raw video from which events will be detected.

There are two threads of process run successively, one runs after the others for a certain time interval t. If web-casting text does not exist, these threads will start from beginning and go through a whole video, otherwise they jump directly to video-event-timestamp pointed out by text-event-timestamp.

1. The former extracts patterns and their timestamp. Those information are recorded in attributes form as $I = (target\_event\_id, event\_id, event\_interval)$ and then saved to attribute table AT.

2. The latter detects events by exploiting the data from AT. The slide-window SW whose length equals *maxlength* is moved along U, each step equals to one camera motion pattern. All patterns occurs inside SW are used to construct a candidate TS γ. Then for every TS s in S, the containment is checked between s and γ. γ will be classified into class S_i if S_i contains one TS α that satisfies:

   - Containment(α, γ) = TRUE
   - if β is the common part of γ and α (i.e all items of β appear both in γ and α), then (1) the length of β must be longest; (2) the different between lengths of γ and α is smallest; and (3) the *confidence* and *support* of β must satisfy predefined thresholds. Note that, γ could have more than one label.

7. EXPERIMENTAL RESULTS

More than 30 hours of soccer video corpus captured at different broadcasters and conditions are used to evaluate the proposed method. Specifically, there are 26 packages of data. Each package contains triplex *(full matches, all events clips extracted from matches offered from broadcaster, web-casting text downloaded from the Internet)*, the second and third item are considered as the ground-truth. We have 20 packages from UEFA champion league, 5 packages from FIFA World Cup 2006, and 20 packages from YouTube (contains only events short clips). We use 10 UEFA, 5 FIFA, and 10 Youtube packages as training set, the rest is used as

testing set. We define 10 events that always appears in all soccer games as follows: *Goal*, *Shot*, *Corner*, *Offside*, *Save*, *Free kick*, *Foul*, *Substitution*, *Red card*, and *Yellow card*.

Figure 3 and 4 illustrate the interface of our system. Figure 3 shows the tasks including in video clock time recognizing, web-casting text parsing, basic events (e.g. pan left, pan right, zoom in, etc) extracting, time-frame aligning, and attributes temporal database constructing. The table in right-bottom corner describes the attributes temporal database D, where each row denotes time when event happen, name of target event, name of basic event, and time interval when that basic event happens. The rich text box in left-right corner shows the content of D under XML format. The rich text box in left-top corner shows the content of web-casting text. The clock time and frame time text box illustrate the time-frame anchor. Figure 4 denotes the NATP mining task. The Event Class list box illustrates the list of target events. The next right list box denotes all temporal patterns mined from attributes temporal database D. The bottom list box shows the number of L-patterns temporal sequences.

Two following evaluation tasks are carried on:

1. ***Quantity***: This evaluation is performed in order to see how many putative events the proposed method could extract from the unknown raw video, and how many events in those putative events are classified into true class. First, a full match video - treated as an input video - is processed by the proposed method to generate a set of all putative events. Then, this set and the set of ground-truth events are used to generate the precision-recall diagram. It should be note that, there is a difference between the case where the input video has web-casting text and where that has not. With the former, since all events are marked exactly by keywords and time stamps that are extracted directly from web-casting text, there is no error or miss in detecting events. In this case, the precision and recall usually equal 100%. Therefore, only the case where there is no support of web-casting text is investigated. Table 3 illustrates the results of the proposed method in the case lack of web-casting text supports.

2. ***Quality***: This evaluation is conducted to see how well the boundary of automatically detected event is. Since up to now, there is no standard condition to check how well the event boundary is, we use the manually labeled events offered by the broadcaster as the system of reference. The Boundary Detection Accuracy (BDA) [20] is use to measure the detected event boundary compared with the ground-truth's.

$$BDA = \frac{\tau_{db} \bigcap \tau_{mb}}{max(\tau_{db}, \tau_{mb})}$$

where τ_{db} and τ_{mb} are the automatically detected event boundary and the manually labeled event boundary, respectively. The higher the BDA score, the better the performance is. Moreover, the method [20] that is used to detect event using web-casting text and Finite State Machine, is also used to compare with our method to distinguish which method is better. Table 4 shows that our method gains the better results than Xu's method.

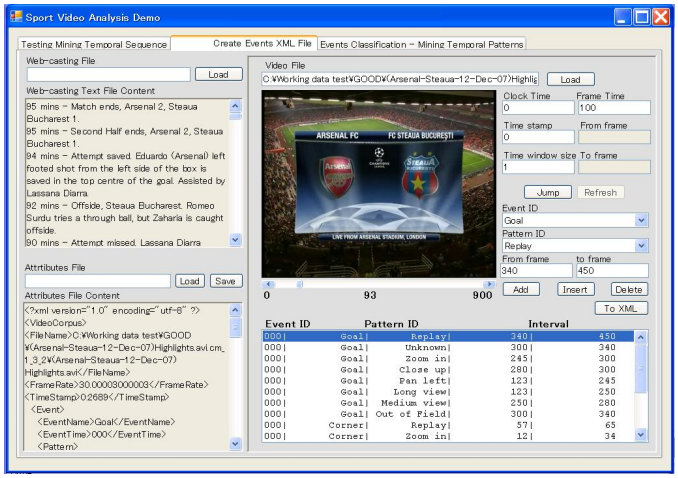


Figure 3: Automatically extracting and annotating the attributes temporal database D from a raw video and web-casting text.

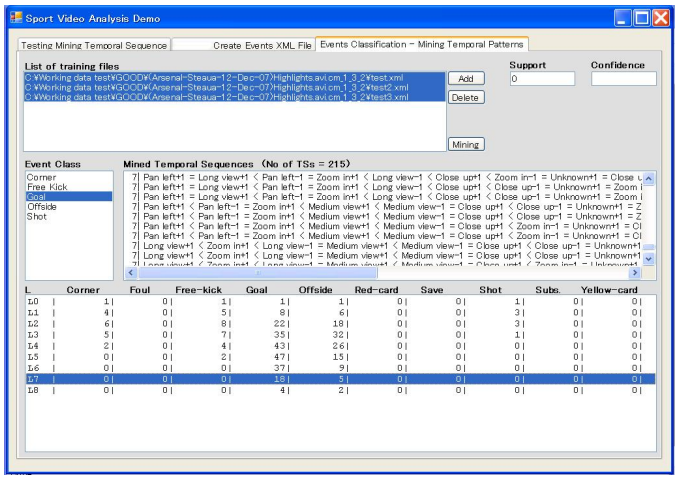


Figure 4: Mining and classifing events using TprefixSpan.

8. CONCLUSIONS AND FUTURE WORK

The new method using non-ambiguous temporal patterns mining and web-casting text to detect event in sports video is presented. Unlike most of existing methods which neglect or use only linear temporal sequence to present temporal information, our method captures temporal information based on Allen temporal algebra. With the support of web-casting text, the training database is generated and annotated automatically. This helps to alleviate the burden of manual annotation process. Moreover, due to the independence between linguistic and perceptual part of patterns, it is easy to deploy this method to another domain (e.g football, baseball, etc.) with only a few modification of perceptual part and transformer. In the future, more features will be considered to find the optimal set of patterns by which the event will be detected more accurate. Beside that, thorough comparisons with related methods will be also conducted to give the better evaluation.

At present, the system offers the text-query scheme in

Table 3: Event detection quantity

Event	Precision/ Recall	Event	Precision/ Recall
Goal	100%/100%	Shot	98%/85.3%
Corner	100%/100%	Offside	100%/100%
Save	100%/100%	Free kick	92%/89%
Foul	83%/80%	Sub.	90%/83.2%
Red card	100%/100%	Yellow card	100%/100%

Table 4: Event detection quality (Pr: the proposed method, Xu: Xu's method [20])

Event	BDA Pr vs Xu	Event	BDA Pr vs Xu
Goal	92% - 76%	Shot	88.2% - 83.1%
Corner	73.1% - 73%	Offside	89.1% - 85.2%
Save	92% - 90.7%	Free kick	44.2% - 43.5%
Foul	81% - 77.7%	Sub.	78% - 78.1%
Red card	83% - 82.5%	Yellow card	84.5% - 84%

which name of players, name of teams, name of stadiums, and name of events which are listed in Table 4 are used as the key-words for querying. In the future, the perceptual part will be paid more attention in order to make the system can be queried by sample (e.g query by a short clip). Moreover, the spatio-temporal information will be investigated instead of dealing with only temporal information in order to detect events more precisely. Beside that the local-sensitive hashing algorithm will also be taken into account to increase the speed of event detection.

9. ACKNOWLEDGMENTS

This research is financially supported by **Japan Society for the Promotion of Science** (JSPS)

10. REFERENCES

[1] J. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, November 1983.

[2] J. Calic, N. Campbell, S. Dasiopoulou, and Y. Kompatsiaris. An overview of multimodal video representation for semantic analysis. In *European Workshop on the Integration of Knowledge, Semantics and Digital Media Technologies, Conference Proceedings*, pages 39–45, December 2005.

[3] M. Chen, S. Chen, and M. Shyu. Hierarchical temporal association mining for video event detection in video databases. In *MDDM'06 Conference Proceedings*, pages 137–145. IEEE, April 2007.

[4] M. Chen, S. Chen, M. Shyu, and K. Wickramaratna. Semantic event edetection via multimoddal data mining. *IEEE Signal Processing Magazine*, pages 38–46, March 2006.

[5] L. Duan, M. Xu, Q. Tian, C. Xu, and J. Jin. A unified framework for semantic shot representation of sports video. *IEEE Trans. on Multimedia*, 7(6):1066–1083, December 2005.

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

[7] M. Fleischman, P. Decamp, and D. Roy. Mining temporal patterns of movement for video content classification. In *MIR'06 Conference Proceedings*, pages 183–191. ACM, October 2006.

[8] M. Fleischman and D. Roy. Unsupervised content-based indexing of sports video. In *MIR'07 Conference Proceedings*, pages 87–94. ACM, September 2007.

[9] S. Jiang, Q. Huang, and W. Gao. Effective image and video mining: an overview of model-based approaches. In *ICME07 Conference Proceedings*, pages 1095–1098. IEEE, July 2007.

[10] Y. Liu, S. Jiang, Q. Ye, W. Gao, and Q. Huang. Playfield detection using adaptive gmm and its application. In *ICASSP'05 Conference Proceedings*, pages 421–424. ACM, March 2005.

[11] R. Missaoui and R. Palenichka. Effective image and video mining: an overview of model-based approaches. In *MDM'05 Conference Proceedings*, pages 43–52. ACM, August 2005.

[12] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu. Mining sequential patterns by pattern-growth: the prefixspan approach. *IEEE Trans. on Knowledge and Data Engineering*, 16(11):1424–1440, November 2004.

[13] D. Sadlier and N. O'Connor. Event detection in field sports video using audio-visual features and a support vector machine. *IEEE Trans. on Circuits and Systems for Video Technology*, 15(10):1225–1233, October 2005.

[14] C. Snoek and M. Worring. Multimedia event-based video indexing using time intervals. *IEEE Trans. on Multimedia*, 7(4):638–647, August 2005.

[15] C. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Journal of Multimedia Tools and Applications*, 35(5):5–34, 2005.

[16] X. Tong, H. Lu, Q. Liu, and H. Jian. Replay detection in broadcasting sports video. In *MIR'05 Conference Proceedings*, pages 337–340. ACM, 2004.

[17] F. Wang, L. Sun, B. Yang, and S. Yang. Fast arc detection algorithm for play field registration in soccer video mining. In *Systems, Man, and Cybernetics Conference Proceedings*, pages 4932–4936. IEEE, October 2006.

[18] S. Wu and Y. Chen. Mining nonambiguous temporal patterns for interval-based events. *IEEE Trans. On Knowledge and Data Engineering*, 19(6):742–758, June 2007.

[19] Z. Xiong, X. Zhou, Q. Tian, Y. Rui, and T. Huang. Semantic retrieval of video. *IEEE Signal Processing Magazine*, pages 18–27, March 2006.

[20] C. Xu, J. Wang, K. Kwan, Y. Li, and L. Duan. Live sports event detection based on broadcast video and web-casting text. In *MM'06 Conference Proceedings*, pages 221–230. ACM, October 2006.

[21] Q. Zhao and S. Bhowmick. Sequential pattern mining: A survey. *ITechnical Report CAIS Nayang Technological University Singapore*, pages 1–26, 2003.

[22] X. Zhu, X. Wu, A. Elmagarmid, Z. Feng, and L. Wu. Video data mining: Semantic indexing and event detection from the association perspective. *IEEE Trans. on. Knowledge and Data Engineering*, 7(5):665–677, May 2005.