A New Linguistic-Perceptual Event Model for Spatio-Temporal Event Detection and Personalized Retrieval of Sports Video

Minh-Son Dao, Sharma Ishan Nath, and Noboru Babaguichi

Media Integrated Communication Lab. (MICL),
Graduate School of Engineering - Osaka University
2-1 Yamadaoka, Suita, Osaka 565-0871, Japan
{dao,sharma}@nanase.comm.eng.osaka-u.ac.jp,
babaguchi@comm.eng.osaka-u.ac.jp

Abstract. This paper proposes a new linguistic-perceptual event model tailoring to spatio-temporal event detection and conceptual-visual personalized retrieval of sports video sequences. The major contributions of the proposed model are hierarchical structure, independence between linguistic and perceptual part, and ability of capturing temporal information of sports events. Thanks to these advanced contributions, it is very easy to upgrade model events from simple to complex levels either by self-studying from inner knowledge or by being taught from plug-in additional knowledge. Thus, the proposed model not only can work well in unwell structured environments but also is able to adapt itself to new domains without the need (or with a few modification) for external reprogramming, re-configuring and re-adjusting. Thorough experimental results demonstrate that events are modeled and detected with high accuracy and automation, and users' expectation of personalized retrieval is highly satisfied.

Keywords: Video signal processing, String matching, Multimedia Information retrieval, Personalization.

1 Introduction

Sports have been known as the most exciting entertainment since the dawn of civilization. People always pay much attention to sports and tend to be involved in active (e.g. players, coaches, etc.) or passive (e.g. audiences, press, etc.) positions as much as possible. Nowadays, with dramatic development of digital technologies and data networks and with unlimited supports of broadcasting industry, audiences' requirements of entertainment have been almost satisfied. Unfortunately, due to different reasons, not all of audiences are ready to spend all their time to watch a full game in a certain sport. In fact, there are only few periods of time that can attract and excite audiences such as "goal" in soccer. These periods of time could be seen as "highlight" or "event".

P. Foggia, C. Sansone, and M. Vento (Eds.): ICIAP 2009, LNCS 5716, pp. 594–603, 2009.
© Springer-Verlag Berlin Heidelberg 2009

Although events could be defined as real-world occurrences that unfold over space and time, many existed methods heavily focus on and exploit the former information (e.g internal-spatial informations)[1]. Although Allen [2] proposed the temporal algebra to model optimally any temporal relation in reality, very little attention has been paid to utilize Allen's theory so far [1][3]. In video analysis, since most of methods dealing with capturing temporal information use only linear temporal relations, they lack the ability to represent temporal information of real complex events compared with Allen's [4][5]. Although there are a few methods that use Allen's algebra to model temporal relations of complex events [6][7], their ambiguous and complex structure representation lead to the increase in computational complexity as well as resources consuming.

In order to deal with one of crucial problems of video analysis: *semantic gap*, most of the existing methods are using supports from the domain knowledge. Since those methods relies heavily on the domain knowledge with significant human interference, it can be hardly applied as a generic framework for an arbitrary domain automatically [1][8].

Moreover, instead of receiving passively whatever products that are offered by broadcasters in one-to-many mode, consumers now tend to request such products that satisfy most of their preferences. The major challenge is the scalability of consumers' preferences. In another word, different users -with different needs and cultures, and accessing the service through heterogeneous terminals and connections- requires different products. Thus, providing tools by which users (both producers and customers) can produce a product based on their own queries related to their preferences seems to be the best solution to satisfy these advanced requirements [9].

In light of these discussions, this paper proposes a new linguistic-perceptual event model tailoring to spatio-temporal event detection and conceptual-visual personalized retrieval of sports video sequences. The major contributions of the proposed model are hierarchical structure, independence between linguistic and perceptual part, and ability of capturing temporal information of sports events. Thanks to these advanced contributions, it is very easy to upgrade model events from simple to complex levels either by self-studying from inner knowledge or by being taught from plug-in additional knowledge. Thus, the proposed model not only can work well in unwell structured environments but also is able to adapt itself to new domains without the need (or with a few modification) for external re-programming, re-configuring and re-adjusting. By taking benefit of the proposed event model, the spatio-temporal event detection method that can adaptively detect events by capturing and representing temporal information using Allen-based temporal algebra is introduced. Results of automatic event detection process are tailored to personalized retrieval via click-and-see style. Thanks to the proposed event model, users could retrieve events by using either conceptual or conceptual-visual fusion query schemes.

2 Linguistic-Perceptual Event Model

Since capturing and presenting temporal information are the core of this research, the temporal database scheme - inspired from Data Mining aspect - is chosen to store events' information. The temporal database contains given events, namely *target_event*, whose contents are the set of patterns that occur in various time periods. These patterns could be other (children) events or concepts, namely *basic_event*. In our method, the temporal database and events are described as follow:

Let $D = \{I\}$ denote the temporal database and

$$I = (target\_event\_name, basic\_event\_name, basic\_event\_interval) \quad (1)$$

denote D's item, where

$$basic\_event\_interval = (time\_start, time\_end) \quad (2)$$

be the time interval where such *basic_event*s happens. The linguistic-perceptual event model is then defined in order to capture informations contained in the temporal database D as follow:

Definition 1 (Event Model)
Given an *target_event* A, the model of A is defined as follow:

$$model(A) \equiv \{(event\_id_i, event\_interval_i)\} \quad (3)$$

$$event\_id_i = (event\_type_i, event\_property_i) \quad (4)$$

$$event\_type_i \in \{codebook\_item_i\} \quad (5)$$

$$event\_property_i \in \{(keyframe\_id_j, cluster\_id_j)\} \quad (6)$$

$$codebook\_item_i = (basic\_event\_name, NATP\_id)_i \quad (7)$$

where *i=1..n*, n is the number of basic events that constructed the target event, *j=1..*m_i, m_i is the number of *cluster_id* that belongs to *event_property*$_i$. The *event_type* is presented as symbols, called *linguistic part*, and the *event_property* captures the event's *perceptual part* that models multimedia patterns (e.g. audio, visual, textual, etc.) of the specializations of the concepts of the linguistic part. The *event_property* is presented by $(keyframe\_centroids, cluster\_id)$ where *cluster_id* is the name of cluster that contains similar keyframes (i.e. similarity distance among these frames is smaller than predefined threshold), and *keyframe_centroids* is the representation of these keyframes. The non-ambiguous temporal patterns (NATP) introduced by Wu et al [3] is applied to create a $NATP\_id$ of the *codebook*. For example: *(camera motion pan left,* $a^{+1} < a^{-1}$*)* is a vector of the codebook.

In other words, each real complex event *target_event* (model(A)) is recorded as the set of *basic_event*s (*event_id*) that occur in complex and varied temporal relations with each other. Each *basic_event* can be low-level feature, mid-level

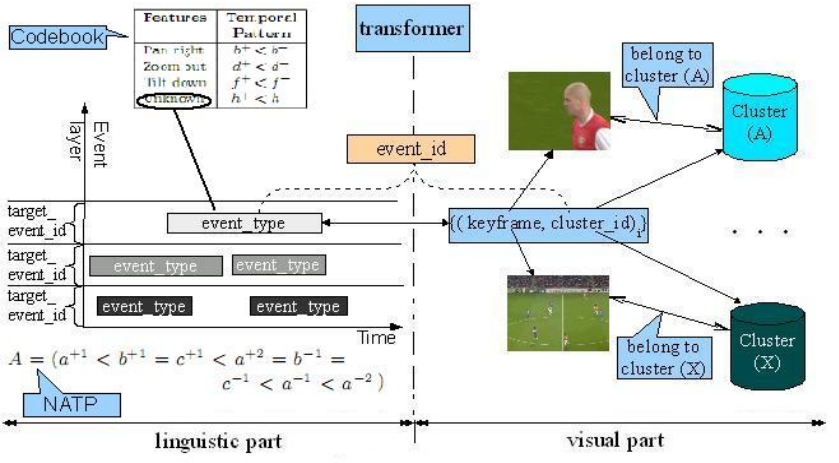

Fig. 1. Illustration of structures of linguistic part, perceptual part (visual concepts), transformer and their relationships of one event

feature, or high-level concept which can be easily extracted from draw video clip using multimedia processing techniques or inferred from simple semantic cues, even a *target_event* of other *basic_event*s

Following sections explain how to apply the proposed event method to spatio-temporal event detection and personalized retrieval. Without loss of generality, in our study, soccer domain is chosen as the case study due to its loose structure of video, the diversification of events, and the highly random occurrence of events.

3 Spatio-Temporal Event Detection

The proposed method has three tasks as follow: (1) create a temporal database D by video-text analysis with supports of web-casting text; (2) represent complex events as temporal sequences, and discover temporal sequences that occur frequently in the temporal database D; and (3) detect events.

3.1 Video and Text Analysis

The purpose of this task is to build the temporal database D from a certain domain. Since this task is already done by authors in the previous works, only compact information which characterized how this task works is presented here. Please refer to [10] for details.

Step 1: Key-words by which events are labeled are extracted from web-casting text by using the commercial software, namely dtSearch[1]. At the end of this step, set of *target_event_id* is built.

Step 2: By linking the time stamp in the text event that is extracted from Web-casting text to the game time in the video that is detected by clock digits

[1] www.dtSearch.com

recognition technique [10], the moment where the event happens is detected. Then, the event boundary is extracted loosely by sliding the time backward and forward from the time-stamp in time interval t.

Step 3:The raw video clip that contains the just-extracted-event is then decomposed into set of simpler (basic) events that were defined beforehand (e.g. the codebook, other event models). When this step finishes, *event_interval* and *event_type* are determined.

Step 4: Key-frames of each basic event are then extracted and clustered into suitable cluster driven by judgment of predefined visual similar measure. *cluster_id* and *keyframe_centroids* are totally defined at the end. Principal Component Analysis and Spectral Clustering are applied for clustering these keyframes.

3.2 Event Presentation and Mining

The mission of this task is to solve the second issue raised in the previous section. In this task, only the conceptual events (i.e. linguistic part) are concerned. Therefore, this task is independent domain-knowledge.

Step 1: First, all *basic_event*s of a certain event are mapped into the codebook to get their temporal patterns presentation. Then, these temporal patterns are aligned by their interval times in the same time axis. Finally, extended Allen temporal algebra is applied to present the certain event as a lexicon of NATP.

Step 2: First, results from the previous step are used to construct the training temporal patterns database (TTPD). Then, the modified NATP framework is used to mine all temporal patterns from TTPD. Since each event is treated as *ruleitem*, mining class association rules is used instead of mining normal association rules to mine temporal patterns (please refer to [10][3] for details).

3.3 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.

The slide-window SW whose length equals *maxlength* is moved along U, each step equals to one camera motion pattern. All patterns occurring 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: (1) Contaiment(α, γ) = TRUE; (2) if β is the common part of γ and α (i.e all items of β appear both in γ and α), then (a) the length of β must be the longest; (b) the difference between lengths of γ and α is the smallest;

and (c) the *confidence* and *support* of β must satisfy predefined thresholds. Note that, γ could have more than one label.

There are two cases for the start position of SW: (1) if web-casting text does not exist, SW will start from beginning and go through a whole video; (2) otherwise SW jumps directly to video-event-timestamp pointed out by text-event-timestamp, then does to-and-fro motion around that timestamp with predefined number of steps.

4 Personalized Retrieval

Thanks to the event model of our method, we can build an event retrieval system that can work well with visual-based and concept-based queries instead of text-based query that is frequently used by most of existing multimedia retrieval systems.

4.1 Query Creating

From the Definition 1, it is not difficult to recognize that *event_id* can be defined by two ways: (1) using only conceptual items that are extracted from the codebook; or (2) using the fusion of conceptual items and visual items that are selected from key-frame database. Therefore, we can provide users two options to construct their own queries as follow: (See figure 1 for visualization of these tasks)

Conceptual query: First, from the codebook (items are showed by text), users select those basic events that - in users' mind - could be a part of an event they want to search. Then, for each *basic_event*, users arrange its interval time on the same time axis. Next, users confirm their query. Finally, the query is generated using the event model that is defined in Definition 1.

Fusion query: First, from the list of $keyframe_id$s of each basic event (items are showed by images), users select their preferred keyframes whose *event_type*s are then automatically mapped using Equation 5. Then, *event_type*s are put on the same time axis. Then, for each *basic_event* w.r.t *event_type*, users arrange its interval time on the same time axis. Next, users confirm their query. Finally, the query is generated using the event model that is defined in Definition 1.

4.2 Re-ranking

The re-ranking task is started when search engine returns results after the first query in order to make the better results in next queries. Set of retrieved results, namely R, are ranked by their confident scores and each result is assigned the ***id*** in its ranking order. Now, users can start re-querying by clicking on any keyframe of any result displayed on monitor as long as a visual content of that keyframe reflects user's imagination of what they want to query.

Let RC denote the set of results users clicked on, $RC = R_i$ where $i \in I$, and I is the set of results' ***id*** where users clicked on. Let S denote the new-query form

constructed by collecting information from users' clicks
$S = \{(event\_id, cluster\_id)\_m\}, m \in I$ so that
$\forall (event\_id, cluster\_id)_m \in S, \exists (event\_id, cluster\_id)_k \in R_i$
so that $(event\_id, cluster\_id)_k = (event\_id, cluster\_id)_m$.

The similarity measure between the new-query S and an arbitrary $C_k \in R$ is defined as
$d(S, C_k) = (\frac{1}{||S||} \sum_{i=1)}^{min(||S||,||C_k||)} (w_1.ES_i + w_2.CS_i)$ where
$ES_i = 1$ if S and C_k have the same $event\_id$, 0 otherwise
$CS_i = 1$ if S and C_k have the same $cluster_id$, 0 otherwise
w_i are weight parameters that satisfy $w_1 + w_2 = 1$, $w_2 > w_1$.

R is then re-ranked according to the value of $d(S, C_k)$. This process is looped until users finish their searching progress.

5 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. 10 students are invited as volunteers. Among them 4 persons are experters, and another 2 persons are naive and the rest are neutral to soccer.

At present, the proposed event model is tested with two levels: (1) *basic_event* level - each *basic_event* is visual or camera motion concept that are defined in the codebook denoted in Table 1; (2) the *target_event* level - we define 10 events that always appear in all soccer games as denoted in Table 2.

Event Detection - 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

Table 1. The Codebook

basic_event_name	NATP_id	basic_event_name	NATP_id
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	$h^+ < h^-$
Long view	$i^+ < i^-$	Medium view	$j^+ < j^-$
Close up	$k^+ < k^-$	Out of field	$l^+ < l^-$
Arc	$m^+ < m^-$	Replay	$n^+ < n^-$
Goal mouth	$o^+ < o^-$	Middle circle	$p^+ < p^-$

Table 2. Event detection evaluation

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

class. 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 2 illustrates the results of the proposed method in the case lack of web-casting text supports.

Event Detection - Quality: This evaluation is conducted to see how well the boundary of automatically detected event is. The Boundary Detection Accuracy (BDA) [4] is used to measure the detected event boundary compared with the ground-truth's. Moreover, the method in [4] that use web-casting text and Finite State Machine, and the method in [6] that utilize the Allen temporal algebra to detect event in sports are also used to compare with our method to distinguish which method is better. Table 3 shows that our method gains the better results than others. It should be noted that Snoek's method focuses on only three events (Goal, Yellowcard, and Substitution)

Personalized Retrieval: All volunteers are asked to query by using both conceptual and fusion queries to retrieve 10 events that are defined in 2 with highest similarity under their own conditions. Table 4 and Table 5 show the volunteers' feedback with respect to conceptual and fusion queries respectively. Most of volunteers satisfy after no more than 5 re-query times. It is easy to see that, naive users can satisfy with retrieval's results within 5 re-query times. Due to the high accuracy of event detection process, the proposed method has the ability to classify an unannotated event into a right class and return to users all data contained

Table 3. Event detection quality (Pr: the proposed method, Xu: Xu's method, Snoek: Snoek's method)

Event	BDA Pr - Xu - Snoek	Event	BDA Pr - Xu - Snoek
Goal	92% - 76% - 86%	Shot	88.2% - 83.1% - N/A
Corner	73.1% - 73% - N/A	Offside	89.1% - 85.2% - N/A
Save	92% - 90.7% - N/A	Free kick	44.2% - 43.5% - N/A
Foul	81% - 77.7% - N/A	Substitution	78% - 78.1% - 78.5%
Red card	83% - 82.5% - N/A	Yellow card	84.5% - 84% - 83%

Table 4. Evaluation of Personalized Retrieval: using Fusion Query and Re-ranking

Trial times	Expert	Naive	Neutral
< 5 times	71%	97%	85%
> 5 times and < 10 times	19%	3%	7%
> 10 times and < 20 times	11%	0%	6%

Table 5. Evaluation of Personalized Retrieval: using Conceptual Query and Re-ranking

Trial Times	Expert	Naive	Neutral
< 5 times	68%	94.2%	83.5%
> 5 times and < 10 times	23%	5.8%	8.3%
> 10 times and < 20 times	9%	0%	8.2%

in that class. Therefore, naive users who are not familiar with soccer, easily accept the retrieval's results without considering the visual similarity such as color of field, color of players' clothes or how the event happens. In contrast, expert users pay more attention in visual similarity so that they need more re-query times.

It is easy to see how query scheme effects the results. When using conceptual query, visual information that is easily recognized by human vision is neglected. This leads to the need of more re-ranking steps to make the final result similar to what users imagine both in conceptual and visual aspects. In contrast, when using fusion query, both conceptual and visual aspects are considered at the beginning. This leads to *near-perfect* result with respect to users' imagination. Thus, less re-ranking steps are needed to get the final result.

6 Conclusions

The new linguistic-perceptual event model is presented. By using the proposed event model, the new generic framework using non-ambiguous temporal patterns mining and web-casting text is built to detect event in sports video tailoring to personalized "click-and-see style" retrieval is presented. Unlike most of existing methods which neglect or use only linear temporal sequence to present temporal information, our method is able to capture and model temporal information of complex event. Moreover, due to the independence between linguistic and perceptual part of patterns, it is easy to deploy this framework to another domain (e.g football, baseball, etc.) with only a few modification of *perceptual part* and *transformer*. Results of automatic event detection progress are tailored to personalized retrieval via click-and-see style. Thanks to our new event model, users could retrieve events by using either conceptual or conceptual-visual fusion query schemes. Moreover, with support of re-ranking scheme, the results after doing query will be pruned to compact final results that are very similar to users' imagination.

In the future, more features and domain will be considered to find the optimal set of patterns by which the events will be detected in high accuracy. Moreover, more higher event levels will be defined to evaluate the ability of self-evolution of the proposed event model. Beside that, thorough comparisons with related methods will also be conducted to give better evaluation.

Acknowledgments

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

References

1. Xie, L., Sundaram, H., Campbell, M.: Event Mining in Multimedia Streams. Proceedings of the IEEE 96(4), 623–647 (2008)
2. Allen, J.: Maintaining knowledge about temporal intervals. Communications of the ACM 26(11), 832–843 (1983)
3. Wu, S., Chen, Y.: Mining nonambiguous temporal patterns for interval-based events. IEEE Trans. on Knowledge and Data Engineering 19(6), 742–758 (2007)
4. Xu, C., Wang, J., Kwan, K., Li, Y., Duan, L.: Live Sports Event Detection Based on Broadcast Video and Web-casting text. In: ACM International Conference on Multimedia, pp. 221–230 (2006)
5. Zhu, X., Wu, X., Elmagarmid, A.K., Feng, Z., Wu, L.: Video Data Mining: Semantic Indexing and Event Detection from the Association Perspective. IEEE Trans. on Knowledge and Data Engineering 7(5), 665–677 (2005)
6. Snoek, C.G.M., Worring, M.: Multimedia Event-Based Video Indexing Using Time Intervals. IEEE Trans. on Multimedia 7(4), 638–647 (2005)
7. Fleischman, M., Roy, D.: Unsupervised Content-based Indexing of Sports Video. In: ACM International Conference on Multimedia Information Retrieval, pp. 87–94 (2007)
8. Xiong, Z., Zhou, X., Tian, Q., Rui, Y., Huang., T.: Semantic retrieval of video. IEEE Signal Processing Magazine, 18–27 (2006)
9. Sebe, N., Tian, Q.: Personalized Multimedia Retrieval: The new trend? In: ACM International Conference on Multimedia Information Retrieval, pp. 299–306 (2007)
10. Dao, M.S., Babaguchi, N.: Mining temporal information and web-casting text for automatic sports event detection. In: International Workshop on Multimedia Signal Processing (2008)