A SEMANTIC EVENT DETECTION APPROACH FOR SOCCER VIDEO BASED ON PERCEPTION CONCEPTS AND FINITE STATE MACHINES

Liang Bai[1,2]*, Songyang Lao*[1]*, Weiming Zhang*[1]*, Gareth J.F.Jones*[2] *, Alan F.Smeaton*[2]

[1]School of Information System & Management, National University of Defense Technology,
ChangSha, China, 410073
bailiang@nudt.edu.cn; laosongyang@vip.sina.com; wmzhang@nudt.edu.cn
[2]Centre for Digital Video Processing, Dublin City University, Glasnevin, Dublin 9, Ireland
{lbai, gjones, asmeaton}@computing.dcu.ie

ABSTRACT

A significant application area for automated video analysis technology is the generation of personalized highlights of sports events. Sports games are always composed of a range of significant events. Automatically detecting these events in a sports video can enable users to interactively select their own highlights. In this paper we propose a semantic event detection approach based on Perception Concepts and Finite State Machines to automatically detect significant events within soccer video. Firstly we define a Perception Concept set for soccer videos based on identifiable feature elements within a soccer video. Secondly we design PC-FSM models to describe semantic events in soccer videos. A particular strength of this approach is that users are able to design their own semantic events and transfer event detection into graph matching. Experimental results based on recorded soccer broadcasts are used to illustrate the potential of this approach.

1. INTRODUCTION

One of the areas of greatest expansion in video content is sports broadcasting. An important component of sports broadcasting is highlights of sports games which are usually prepared manually. However, these are not always available and the material included is selected by a single editor. This situation is inflexible with respect to individual viewers who may want a longer or shorter summary or to focus on certain event types; and the need for manual editing means generation of summaries is often not cost effective.

Video technology can potentially provide new ways to view soccer videos which are more interactive and personal, rather than to view passively pre-edited highlights (when they are actually available). It is important for video processing and retrieval researchers to develop new ways for users to search for semantic events within soccer videos. To date work in this area has focused on query-by-text. The user enters the word “goal”, then the system searches for video clips including the word goal in the soundtrack or possibly in manually added metadata [1][2]. This clearly relies on either a well annotated soundtrack or a costly manual labeling of semantic events. Automatic detection of semantic events that captures the essential contents of a game is becoming more and more important. Related prior work towards automatic events detection in sports videos is described in [3][4][5]and[6]. In most existing work the event detection algorithms are embedded in systems and cannot easily be redefined. This means that users cannot adapt the event types detected or the system refined to the different editorial rules used by different broadcasting corporations. A semantic description method based on Petri-Net is described in [7].Through experiments we found that Petri-Net it is complex to generate SQL queries. In this work relatively simple graph model, FSM, is used. And BSU defined in [7] give us an inspiration to define Perception concepts (PCs).

In this paper we introduce a semantic event detection approach for soccer video. For this approach we first define perception concepts to describe patterns sharing similar spatio-temporal behaviors in soccer video. PCs are combined in Finite State Machines which describe the spatio-temporal relations between the PCs associated with significant events in a soccer game. PC-FSMs are described formally in term of state graphs. A graph matching method is used to detect semantic events automatically. Finally we illustrate the validity of this model using experiments on recorded soccer videos.

2. PERCEPTION CONCEPTS IN SOCCER VIDEO

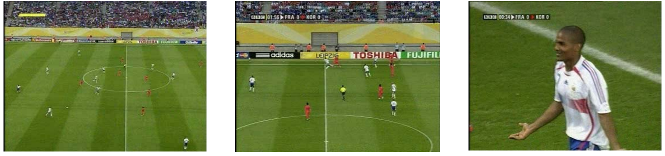
In soccer games program editors are interested in selecting similar and periodical action which can help the audience to understand and enjoy the game. In this case, it is important that the similar and periodical actions patterns share similar spatio-temporal behaviors that can be clustered and described with a linguistic concept. These requirements motivate the possibility that patterns that share the same behaviors can be represented by PCs. PCs in soccer video are abstractions of video elements and can be of two main types: Visual Concepts and Aural Concepts. In this section we outline the characteristics of PC types.

- Visual Concepts

Visual concepts in sports videos share the same visual features. Visual concepts can be of different types: sequence, object and slow-motion-replay which are a special and important component of soccer video.

Eight International Workshop on Image Analysis for Multimedia Interactive Services(WIAMIS'07)
0-7695-2818-X/07 $20.00 © 2007 IEEE

Sequence: According to the focus of cameras, Sequence Visual concepts can be classified as: Loose View, Medium View and Tight View (see Figure 1).

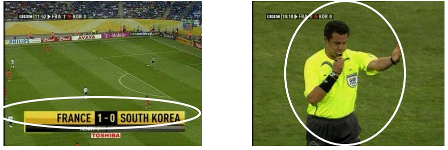


(a) Loose View (b) Media View (c) Tight View

Figure 1 Sequences in Soccer Game

Dominant color feature which represents local features where a small number of colors are enough to characterize the color information in the region of interest can be used for sequence classification [8]. The loose view and medium view share analogical visual features and are often associated with one shot zoom action. So they can be defined as one visual concept style named *Normal View (NV)* in this paper. When some highlights occur, the camera view often focuses on a normal view to capture something interesting in the auditoria. So the normal view in the case of out-of-field is an important visual concept.

Object: Only a limited number of object types are observed in sports videos, such as: ball, player, referee (and assistant referees), coach, captions and so on. In this paper we only select two types of object: caption and referee (see Figure 2). These can be detected reliably in soccer video [8], and are thus available to be described in semantic content. We illustrate later that these object PCs can be used effectively in the description of significant events in soccer video.



(a) Caption (b) Referee

Figure 2. Objects

Slow-motion-replay: In soccer videos important semantic events are often replayed in slow motion immediately after they occur. The pixel-wise mean square difference of the intensity of every two subsequent frames and RGB color histogram of each frame can be used in a HMM model for slow-motion-replay detection [9].

- Aural Concept

Aural concepts are useful for semantic analysis of games. In general, in a soccer match there are two kinds of important audio: whistle and cheers. High crowd noise with low or absent speech often means the cheers. A whistle from a referee has high frequency and a strong spectrum [10]. It can be detected according to peak frequencies which fall within the threshold range. A whistle sounds useful to suggest that something interesting would happen. So we consider the two types of aural concepts in this paper.

3. DEFINITION OF PC-FSM MODEL

FSM is an abstract machine consisting of a set of states (including the initial state), a set of input events, a set of output events, and a state transition function. The function takes the current state and an input event and returns the new set of output events and the next state. FSM is effective for modeling sequential process. Formally, a PC-FSM is defined as follows:

Definition 1 A PC-FSM is a 7-tuple

$$C_{PC-FSM} = \{S_{PC}, S_0, I, O, T, Op, Dt\}$$

where:

S_{pc} is the set of states in sports video.

$$S_{pc} = \{IF\_NV, OF\_NV, TV, SMR\}$$

The elements in S_{pc} respectively represent *Infield Normal View*, *Out of field Normal View*, *Tight View*, *Slow Motion Replay*.

S_0 represents the initial state. In soccer games when an event begins and ends the camera view mainly focuses on the field with an "*infield loose view*" of the action. So we can set S_0 as *infield loose view.*

I is the input event set. In soccer videos the camera focus will be changed when some events happen, such as a foul. For the model, this drives the transitions of PC states. For a strict description, *Null* describes transition without any input events and *Event-End* indicates the end of a semantic event.

O is the output set. When an *I* event happens, some perception concepts occur, such as whistle or caption.

T is a finite set of transitions. Each of the transitions *t* in T can be defined as follow:

$$t : \langle Head(t), I(t), O(t, op), Tail(t) \rangle$$

where, *Head (t)* is the starting state; *I (t)* is the input event of *t*; *O (t,op)* is the output event set of *t*; *op* indicates the logic relation between output events and *t*, it can be figured as: $Event \mid op, op \in Op$. $Tail(t)$ is the ending state.

Op is a set of logic operators that indicate the logic of relations among events and between events and transition. For the PC-FSM model we define *Op* set as follow:

$$Op = \{following, before, synchronazition\}$$

We define before, following and synchronization as tokens to describe the sequence of output events and transitions. *Dt* is defined as the duration time of a state or an output event.

Definition 2 *Dt* is a 2-tuple

$$Dt(Operator, Time).$$

where $Operator = \{\le, =, \ge\}$, $Time \in R$.

Dt is an important parameter in the query processing. For example, in the semantic of Goal, the duration of *Tight View* state is very long.

In PC-FSM model, the set of output events O is as follows:

$$O = \{Null, Visual\ Objects, Aural\ Concepts\}$$

where:

$Visual\ Object = \{Caption, Corner\ Arc, Referee\}$

$Aural = \{Cheer, Whistle\}$. The *Null* element represents no output events occurring during a transition.

4. SEMANTIC EVENTS DESCRIPTION AND DETECTION BASED ON PC-FSM MODEL

A soccer game is mainly composed of In-Play and Out-of-Play. Some events always bring on Out-of-Play, such as a foul or throw-in. The audience is often interested in In-Play and the Events. So we classify the semantic content of soccer video into: In-Play Scene and Events. We selected five matches from the 2006 World Cup to design PC-FSM models for different semantic events. A particular strength of this approach is that the user can modify or define new PC-FSMs to describe soccer semantic events based on their knowledge of activities in soccer matches.

4.1 In-Play Scene

When the ball is in play the camera view mainly focuses on the field with an “Infield Medium” or “Infield Long” view of the action, and a few tight views internally. The PC-FSM of In-Play Scene is shown in Figure 3.

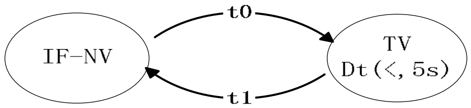


Figure 3. PC-FSM Description Graph of In-Play Scene

Transitions in Figure 3 are defined as follow:

$t0:\langle IF-NV, Null, Null, TV\rangle$

$t1:\langle TV, Null, Null, IF-NV\rangle$

So if no event happens, the ball is always in play and the PCs states transit between IF-NV and TV.

4.2 Events

In this paper, the events described are: Goal Scored and Foul, but other events can easily be described using the same method.

- **Goal Scored**

The PC-FSM of Goal Scored is shown in Figure 4.

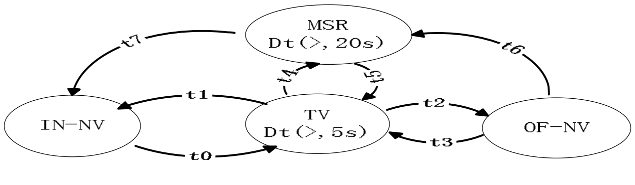


Figure 4. PC-FSM Description Graph of Goal Scored

Transitions in Figure 4 are defined as follow:

$t0:\langle IF-NV, Goal\ Scored, \{Whistle \mid before, Cheer \mid synchronization\}, TV\rangle$

$t1:\langle TV, Event-End, \{Caption \mid following\}, IN-NV\rangle$

$t2:\langle TV, Null, \{Cheer \mid synchronization\}, OF-NV\rangle$

$t3:\langle OF-NV, Null, \{Cheer \mid synchronization\}, TV\rangle$

$t4:\langle TV, Null, \{Cheer \mid synchronization\}, MSR\rangle$

$t5:\langle MSR, Null, \{Cheer \mid synchronization\}, TV\rangle$

$t6:\langle OF-NV, Null, \{Cheer \mid synchronization\}, MSR\rangle$

$t7:\langle MSR, Event-End, \{Caption / following\}, IF-NV\rangle$

- **Foul**

The PC-FSM of Foul is shown in Figure 5.

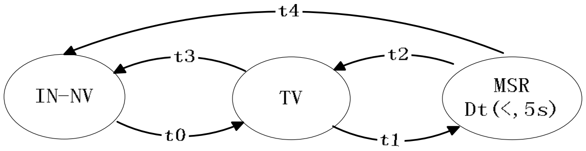


Figure 5. PC-FSM Description Graph of Foul

Transitions in Figure 5 are defined as follow:

$t0:\langle IF-NV, Foul, \{Whistle \mid before, Referee \mid following\}, TV\rangle$

$t1:\langle TV, Null, Null, MSR\rangle$

$t2:\langle MSR, Null, Null, TV\rangle$

$t3:\langle TV, Event-End, Null, IN-NV\rangle$

$t4:\langle MSR, Event-End, Null, IN-NV\rangle$

If the foul causes a yellow card or a red card, there will be some difference for PC-FSM model:

TV state and MSR state will last longer time:

$$TV: Dt(>, 10s)\ ;\ MSR: Dt(>, 10s)$$

Transition t0 will be an output Caption:

$$t0:\left\langle IF-NV, Foul, \begin{Bmatrix} Whistle \mid before, Referee \mid following, \\ Caption \mid following \end{Bmatrix}, TV\right\rangle$$

4.3 Semantic Events Detection

Based on the PC-FSM model, automatic events detection in soccer video can be designed. For the matches videos, shot cluster method or manual annotation can be used to annotate PC state for each video shot. The experiments described here used a manually annotated database of video shots. This was used in order to eliminate effects of errors in shot cluster, since in this paper we focus on demonstrating the validity of the PC-FSM model. After manual annotation the video of a match is converted to states with output PCs. The events detection is carried out in two steps.

First Step: candidate events are detected in sequence of video states. The target video is segmented by occurrence of IF-NV state. The segmentation algorithm is as follow:

```
Initially: i = 0, j = 0;
do {
        if CurrentShot.StateType == IF-NV
        {
            if i==0
                i = CurrentShot.ShotID;
            else
            {
                j = CurrentShot.ShotID;
                Create Segmentation(i, j);
                Segmentation(i, j).StartTime = i.StartTime;
                Segmentation(i, j).EndTime = j.EndTime;
                i = j; j = 0;
            }
        }
        Update CurrentShot with the next shot;
    } while (CurrentShot Exist);
```

If segmentation is without any output events, it is annotated with In-Play Scene. Other Out-of-Play Scene segmentations are considered as candidate events in which the semantic events perhaps occur.

Second step: is matching between candidate events and PC-FSM model of semantic content, and detecting the semantic events. The candidate event can be described formally as a state graph like PC-FSM state graph. Then events detection can be carried out using graph matching method.

Definition 3 For PC-FSM graphs G_a and G_b, if $S_a = S_b$, $Dt_a = Dt_b, for\ each\ S$ and $transitionSet_a \subseteq transitionSet_b$, then $G_a \subseteq G_b$.

The rule for semantic events detection is: if the PC-FSM graph for a candidate event belongs to a given PC-FSM graph defined section 4.1 and 4.2, a semantic event are detected and annotated. For example: if $G_{candidate} \subseteq G_{goal-scored}$, Candidate event is goal-scored.

5. EXPERIMENTS AND EVALUATION

In order to demonstrate our approach to identifying semantic content in sports video we conducted a preliminary set of experiments. These were carried out using five soccer games recording captured from 4:2:2 YUV PAL tapes which are saved as MPEG1 format. The soccer videos are from a range of broadcasters (ITV and BBC Sport), and are taken from the 2006 World Cup, and are 7hs 53mins28s long. The PC-FSM models were developed initially using one game with some subsequent minor adjustments based on the other four. Table 1 shows the time and percentage of each PC state type in the five matches.

Table 1. The duration and PC state percentage

ID	PC State Name	Percentage of Duration	Percentage of PC state Number
0	IF-NV	54.75%	40.28%
1	TV	33.58%	46.17%
2	MSR	9.41%	9.23%
3	OF-NV	2.26%	4.32%

Table 2 shows “Precision” and “Recall” for detection of the semantic events. “Actual Num” is the actual number of events in whole matches; “True Num” is the number of detected correct matches, and “False Num” is the number of false matches.

Table 2. Precision and recall for five soccer semantics

semantic	Actual Num	True Num	False Num	Precision (%)	Recall (%)
In-Play Scene	486	447	79	92.0%	71.5%
Goal Scored	10	10	0	100%	100%
Foul	193	169	19	89.9%	87.6%
Yellow (or Red) Card	26	26	2	92.9%	100%

From Table 2, it can be seen that the results of events detect are higher than 87% and the recall of In-Play Scene detecting is 92.0%. The precision of In-Play Scene detecting is relatively low. Because state transitions of Throw-In event is similar with In-Play. Some players often kick off quickly after a foul, so TV and MSR states will not occur. This is the reason for losing some true foul events. When a player is injured, TV lasts for a long time and MSR and Caption object will occur. In this case, a yellow card event is decided wrongly.

Based on the above experimental results, we believe that this approach to searching in sports video has considerable potential. We are currently conducting a more thorough experimental investigation using a larger set of independent videos.

6. CONCLUSIONS AND DISCUSSIONS

In this paper, based on analyzing of sports video characteristics, we define a Perception Concepts for soccer games and proposed a semantic events detection approach using Finite State machines. The effectiveness of the proposed approach was demonstrated through a preliminary experiment. The approach can be utilized for different sports video. Future work will explore interface design, which is very important to enable rapid development of new queries demands, and research intelligent methods for PC-FSM graph matching.

7. REFERENCES

[1] D. Zhang and D. Ellis. Detecting Sound Events in Basketball Video Archive. Technical Report, Dept. of Electrical Engineering, Columbia University, 2001.

[2] R. Dahyot, A. C. Kokaram, N. Rea and H. Denman. Joint Audio-Visual Retrieval for Tennis Broadcasts. In Proceedings of the 28th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’03), Hong Kong, April 2003.

[3] Dongqing Zhang, Shih-Fu Chang. Event Detection in Baseball Video Using Superimposed Caption Recognition. In Proceedings of the 10th ACM International Conference on Multimedia, pp315-318, Juan-les-Pins, France, Dec 2002.

[4] Riccardo Leonardi and Pierangelo Migliorati. Semantic Indexing of Multimedia Documents. IEEE MultiMedia, Vol.9, No. 2, pp44-51, April/June 2003.

[5] Wensheng Zhou, Asha Vellaikal, C. C. Jay Kuo. Rule-based Video Classification System for Basketball Video Indexing. In Proceedings of the 8th ACM International Conference on Multimedia, pp213-216, Los Angeles, CA, USA, Oct 30-Nov 04, 2000.

[6] S. Nepal, U. Srinvasan, G. Reynolds. Automatic Detection of 'Goal' Segments in Basketball Videos. In Proceedings of ACM Multimedia '01, Canada, 2001.

[7] Songyang Lao, Alan F. Smeaton, Gareth J. F. Jones, Hyowon Lee. A Query Description Model Based on Basic Semantic Unit Composite Petri-Nets for Soccer Video Analysis. In Proceedings of ACM MIR’04, October 15–16, 2004, New York, USA

[8] J.Y. Chen, Y.H. Li, S.Y. Lao, et al, Detection of Scoring Event in Soccer Video for Highlight Generation. Technical Report, National University of Defense Technology, 2004.

[9] Hao Pan, P. van Beek, M. I. Sezan. Detection of Slow-motion Replay Segments in Sports Video for Highlights Generation. In Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP’01), Salt Lake City, UT, USA, May 2001.

[10] Zhou, W., S. Dao, and C.-C. Jay Kuo, On-line knowledge and rule-based video classification system for video indexing and dissemination. Information Systems, 2002. 27(8): p. 559-586.

0-7695-2818-X/07 $20.00 © 2007 IEEE