A SEMANTIC CONTENT ANALYSIS MODEL FOR SPORTS VIDEO BASED ON PERCEPTION CONCEPTS AND FINITE STATE MACHINES

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

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

ABSTRACT

In automatic video content analysis domain, the key challenges are how to recognize important objects and how to model the spatiotemporal relationships between them. In this paper we propose a semantic content analysis model based on Perception Concepts (PCs) and Finite State Machines (FSMs) to automatically describe and detect significant semantic content within sports video. PCs are defined to represent important semantic patterns for sports videos based on identifiable feature elements. PC-FSM models are designed to describe spatiotemporal relationships between PCs. And graph matching method is used to detect high-level semantic automatically. A particular strength of this approach is that users are able to design their own highlights and transfer the detection problem into a graph matching problem. Experimental results are used to illustrate the potential of this approach.

1. INTRODUCTION

Nowadays we can access increasing amounts of video data as a result of the development of high-speed broadband networks and digital video technology. 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. Preparing this is time-consuming and the situation is inflexible with respect to individual viewers who may want a longer or shorter summary or to focus on certain event types. Further the need for manual editing means that often the generation of any summary at all is not cost effective.

It is important for video processing and retrieval researchers to develop new ways to model interesting semantic content and identify semantic events automatically in sports 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 discovery of semantic content that captures the essential contents of a game is becoming more and more important. Related prior work towards automatic event 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. In our work, we have proposed a semantic description method based on Petri-Net in [7] and used Finite State Machines to detect semantic events in soccer video [8]. In this paper we introduce a semantic content analysis model for sports video. For this model we first define PCs to describe patterns sharing similar spatiotemporal behaviors in sports video. PCs are combined into FSMs which describe the spatiotemporal relations between the PCs associated with significant semantic content in a game. PC-FSMs are described formally in terms of state graphs. A graph matching method is used to discover semantic content automatically. Finally we illustrate the validity of this model using experiments on recorded sports videos.

2. PERCEPTION CONCEPTS IN SPORTS VIDEO

In the sports domain, a TV sports program editor is interested in selecting similar and periodic actions which can help the audience to understand and enjoy a game. In this case, it is important that the similar and periodic-action patterns share similar spatiotemporal 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 sports 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.

Sequence: a sequence of frames captured by one camera in a single continuous action with fixed focus and steady content.

Object: a segmentable region representing a distinct semantic concept in a frame.

Slow-motion-replay (SMR): is a sequence of frames played in a low play rate to replay video content of an earlier event, which is a special and important component of sports video.

- Aural Concept

An aural concept is a temporal segmentation of the audio track and represents a distinct semantic audio type. Aural concepts are generated by audio segmentation and classification. Aural concepts, such as: cheers from the audience and speech from commentators, are useful for semantic analysis of games.

3. DEFINITION OF PC-FSM MODEL

A FSM is an abstract machine consisting of a set of states, 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. FSMs are effective for modeling sequential processes. 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.

I is the input event set. For a PC-FSM 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 a special object.

T is a finite set of transitions. Each of the transitions t in T can be defined as : $t: \langle Head(t), I(t), O(t,op), Tail(t) \rangle$

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 end state.

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

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

Video is different to text since there are temporal relations between events and transition. So we define following, before and synchronization as tokens to describe the sequence of output events and transitions.

Dt represents the duration time of a state or an output event.

Definition 2 Dt is a 2-tuple

$$Dt(Operator, Time).$$

where: $Operator = \{\leq, =, \geq\}$, $Time \in R$.

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:

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

The *Null* element represents no output events occurring during a transition.

A PC-FSM can be represented by a directed graph $G(V, E)$ institutively, where V, the vertexes set, represents the states of PC-FSM and E, the edges set, represents the transitions between different states.

4. SEMANTIC CONTENT DESCRIPTION AND DISCOVERING BASED ON PC-FSM MODEL

Usually, a sports match is composed of In-Play and Out-of-Play. Out-of-Play events are always brought about by special events, such as a foul in soccer or basketball. An In-Play scene is composed of a mass of LVs and MVs with the playing field as background and a spot of TVs. An Out-of-Play scene is composed of TVs, LVs with auditoria as background and MSRs inserted by the editor. Events in sports matchs often occur in In-Play and drive a sequence of Out-of-Play events, including TV interviews with relevant people, such as players or referees, and MSRs for replay for the events and so on. This structure of sports video content can be described as in Figure 1.

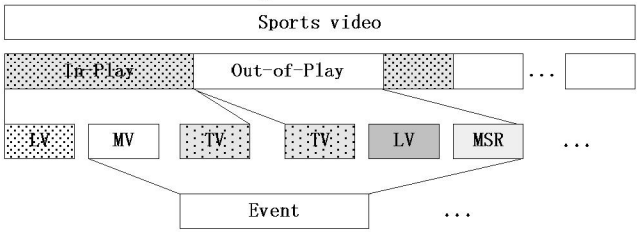

Figure 1. Structure of sports video content

This structure can be studied from a training video data set and defined by a PC-FSM model according to the user's knowledge of a sports domain. Then we can automatically discover semantic content using defined PC-FSM models. We take popular sports, soccer and basketball, as examples for the application of the PC-FSM model.

4.1 Semantic Content Description

According to observations for sports video, sequence visual concepts can be classified as: Out-of-Field View (OFV), In-Field Loose View (IF-LV), In-Field Medium View (IF-MV) and Tight View (TV) (see Figure 2). 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 (for example green for soccer field and yellow for basketball in Figure 2), can be used for sequence

classification [9]. The IF-LV and IF-MV share analogical visual features and are often associated with a one-shot zoom action. They can be defined as one visual concept style named In-Field Normal View (IN-NV), as adopted in this paper.

Loose View Media View Tight View OFV

Figure 2 Sequences in Sports Game

.Only a limited number of object types are observed in sports videos, such as: player, referee, coach, captions and so on. Here we only select one object: caption, which can be detected reliably in sports video [9], and is thus available to be described as semantic content. In general, in a sports match there are two kinds of important audio: whistles and cheers. High crowd noise with low or absent speech is often associated with loud 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 sound is useful to suggest that something interesting has happened. So we consider whistle and cheers as aural concept here. We describe Goal Scored in soccer and Foul in basketball using PC-FSM model. S_0 is set to IN-NV because when an event begins and ends the camera view mainly focuses on the field with an IN-NV.

- **Goal Scored in soccer**

The PC-FSM of Goal Scored is shown in Figure 3.

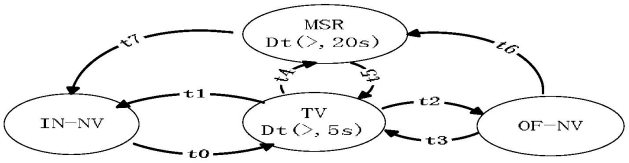

Figure 3. PC-FSM Graph of Goal Scored

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

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

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

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

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

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

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

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

- **Foul in basketball**

The PC-FSM of Foul is shown in Figure 4.

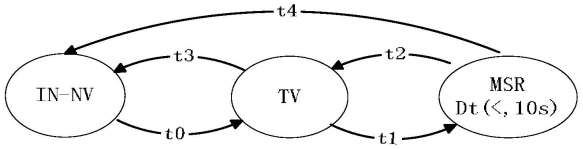

Figure 4. PC-FSM Graph of Foul

Transitions in Figure 4 are defined as follow:

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

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

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

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

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

Other semantic content can be similarly described in the same way using the PC-FSM model. A particular strength of this approach is that the user can modify or define new PC-FSMs to describe different semantic content based on their knowledge of activities in the sports domain.

4.2 Semantic Content Discovering

Based on the PC-FSM model, automatic discovery of semantic content in sports video can be designed. 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 and object recognition, 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 a series of states with output PCs. Semantic content discovery is carried out in two steps:

Firstly, candidate events are detected in the sequence of video states. According to the definition of the PC-FSM model, the initial state is IF-NV. So the sports video is segmented by occurrence of the IF-NV state. The sequence between one IF-NV and the next is considered as a candidate event.

The second step is matching between candidate events and the PC-FSM model of semantic content. The candidate event can be described formally as a state graph with some transitions like PC-FSMs graph. Then event detection can be carried out using a 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 event detection is: if the PC-FSM graph for a candidate event belongs to a given PC-FSM graph, a semantic event is detected and annotated.

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 and three basketball games recorded from 4:2:2 YUV PAL tapes as MPEG-1 format. The soccer videos are from two UK broadcasters (ITV and BBC Sport), and are taken

from the 2006 World Cup, taking a total of 7hs 53mins 28s. And the basketball videos are NBA games recorded from ESPN, FOX Sports and CCTV5 taking a total of 6hs 47mins 18s.

For soccer videos we defined Goal Scored (GS), Foul (S-F) and Yellow (or Red) Card (YRC) events and detect them based on PC-FSM model. Highlight Attack (HA) and Foul (B-F) events are defined and detected in basketball videos. Table 1 shows "Precision" and "Recall" for detection of the semantic events. The ground truth is recognized manually. It can be seen that the precision and recall are higher than 87%.

Table 1. Precision and recall for sports semantics

semantic	GS	S-F	YRC	HA	B-F
Pre (%)	100	89.9	92.9	88.4	87.7
Rec (%)	100	87.6	100	93.2	96.5

We also compared our approach with other approaches. In our previous work, a Petri-Net (PN) model is used for video semantic content description and detection [7]. HMM is a popular model for video event detection [5][9]. In our experiments, we use the PN based approach and HMM based approach proposed in [9] to detect semantic content using same video data set. The results are shown in Table 2.

From Table 2, we can find the precision and recall of PN based approach is almost equivalent with PC-FSM based approach. But the key problem for PN based approach is difficult to transfer a complex PN model to SQL query automatically. PN model is stronger in semantic content description but weaker in detection processing.

Low precision and recall are shown in the experimental results of HMM based approach, in which low-level features, color, texture and motion vector, are extracted to training different HMM models for different semantic content. This approach tends to map low-level features to high-level semantic directly according to statistic point, which can capture perception feature pattern well but not be effective to model and detect spatiotemporal relationship between different semantic content.

Table 2. Results based on PN and HMM Approach

semantic		GS	S-F	YRC	HA	B-F
PN	**Pre(%)**	85.2	86.6	91.7	85.8	84.5
	Rec(%)	100	84.1	97.5	91.6	90.3
HMM	**Pre(%)**	75.4	63.8	77.6	61.5	59.2
	Rec(%)	80.1	72.5	83.1	64.9	67.3

The PC-FSM based approach defines PCs as middle-level semantic content that represents perception patterns and narrows the gap between low-level features and high-level semantic content. Using a graph match method for discovering high-level semantic content avoids the difficulty of generating complex SQL queries. Based on the above experimental results, we believe that our approach to searching in sports video has considerable potential.

6. CONCLUSIONS AND DISCUSSIONS

In this paper, based on analyzing sports video characteristics, we define PCs for sports games and proposed a semantic content analysis model using FSMs. The effectiveness of this model was demonstrated through preliminary experiments. Our method can be utilized for different sports video. Future work will explore interface design, which is very important to enable rapid development of descriptions for the new event types, and research intelligent methods for PC-FSM model graph matching.

7. ACKNOWLEDGMENTS

This work is supported by the National High Technology Development 863 Program of China (2006AA01Z316), the National Natural Science Foundation of China (60572137) and China Scholarship Council of China Education Ministry.

8. 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] D. Zhang, S-F. 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] R.Leonardi and P. Migliorati, "Semantic Indexing of Multimedia Documents", IEEE MultiMedia, Vol.9, No. 2, pp44-51, April/June 2003.

[5] Guoying Jin, Linmi Tao, Guangyou Xu, "Hidden Markov Model Based Events Detection in Soccer Video", International Conference of Image Analysis and Recognition, Porto, Portugal, October 2004, LNCS 3211, pp. 605-612

[6] D.A.Sadlier and N.E. O'Connor, "Event Detection in Field Sports Video Using Audio-visual Features and A SVM", IEEE Transactions on Circuits and Systems for Video Technology, 15(10), 1225-1233, 2005.

[7] S.Y. Lao, A. F. Smeaton, G. J. F. Jones, H. 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] L. Bai, S.Y. Lao, W.M. Zhang, A. F. Smeaton, G. J. F. Jones, "A Semantic Event Detection Approach for Soccer Video based on Perception Concepts and Finite State Machines", International Workshop on Image Analysis for Multimedia Interactive Services, Santorini, Greece, 6-8 June 2007.

[9] 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.

[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.