Semantic Analysis of Field Sports Video using a Petri-Net of Audio-Visual Concepts

Liang Bai[1,2], Songyang Lao[1], Alan F. Smeaton[2], Noel E. O'Connor[2], David Sadlier[2], David Sinclair[3]

[1]*School of Information System and Management, National University of Defense Technology, ChangSha, China, 410073,* [2]*Centre for Digital Video Processing and Adaptive Information Cluster, Dublin City University, Glasnevin, Dublin 9, Ireland.* [3]*Lero, School of Computing, Dublin City University, Glasnevin, Dublin 9, Ireland.*
Email: lbai@computing.dcu.ie

The most common approach to automatic summarisation and highlight detection in sports video is to train an automatic classifier to detect semantic highlights based on occurrences of low-level features such as action replays, excited commentators or changes in a scoreboard. We propose an alternative approach based on the detection of perception concepts (PCs) and the construction of Petri-Nets which can be used for both semantic description and event detection within sports videos. Low-level algorithms for the detection of perception concepts using visual, aural and motion characteristics are proposed, and a series of Petri-Nets composed of perception concepts is formally defined to describe video content. We call this a Perception Concept Network-Petri Net (PCN-PN) model. Using PCN-PNs, personalized high-level semantic descriptions of video highlights can be facilitated and queries on high-level semantics can be achieved. A particular strength of this framework is that we can easily build semantic detectors based on PCN-PNs to search within sports videos and locate interesting events. Experimental results based on recorded sports video data across three types of sports games (soccer, basketball and rugby), and each from multiple broadcasters, are used to illustrate the potential of this framework.

Received October 2007; revised June 2008

1. INTRODUCTION

One of the areas of most significant growth in video content is in sports broadcasting where ever greater numbers of events are broadcast live and in their entirety. It is not possible for even the most avid sports fan to watch more than a small fraction of the available coverage. Furthermore, much of the coverage is often not significant to the progression of the game or its outcome. For this reason, presentation of highlights of sports games is extremely important for broadcasters and viewers alike, but typically a manual, time-intensive process is required to prepare these highlights. Ideally it should be possible to personalize these highlights for viewers who may want longer/shorter highlights, but this is not easily facilitated by a manual system. Thus, an important practical requirement for real systems is to describe and detect high-level semantic content and to generate personalized highlights automatically. In this paper, we propose a novel semantic content analysis framework for sports video based on Perception Concepts Net and Petri-Net, called PCN-PN, that addresses these requirements. We believe that the Petri-Net formalism is ideally suited to capturing the temporal and sequential relationships among occurrences of basic perception concepts in sports video highlights while at the same time being able to cater for temporal re-arrangement of these concepts, where a sports broadcaster may vary the sequence of these concepts. Petri-Nets allow flexible encoding of this variability while depending upon accurate detection of the lower-level concepts.

In the literature, the work on sports video processing can be broadly divided into the identification of mid-level semantic content such as objects, the detection of high-level semantic content such as events or scenes, and then highlight generation (summarisation). Most

approaches are based upon an initial low-level aural, visual and motion feature extraction stage. Specifically, audio feature extraction has been shown to be useful in assisting high-level semantic content detection across a wide variety of sports [1, 2, 3]. Dominant color can be employed to classify shots in certain genres [1, 4]. Object motion trajectories and interactions have been used for football play classification [5] and for soccer event detection [6]. Object identification clearly also provides important cues for high-level semantic content detection, such as shot classification [7], player identification [8,9], player action recognition [10,11], referee detection [4,12], ball tracking [12,13], caption detection [1,14], and slow motion replay detection [15]. According to results reported in the literature, the best performances are achieved by using multimodal features.

The rest of this paper is organised as follows, In the next section we give some background and related work and following that we present the visual, aural and motion concepts that we use in our work. Some low-level video processing algorithms for audio and visual concept detection are then proposed in section 4. Section 5 presents the formal definition of the Perception Concepts Net & Petri-Net model (PCN-PN), and we elaborate on this by showing different levels of semantic content description in three types of field ball game videos using PCN-PNs. In Section 6 we present experimental results on perception concept detection and high-level semantic event detection over more than 15 hours of video data obtained from multiple broadcast sources and then we conclude the paper.

2. BACKGROUND

Due to dramatic variances in broadcast styles for different sports, much of the prior art in this area tends to target one type of sport. Soccer video is by far the most popular [4, 16, 17, 18]. However, audio and/or video based analysis for high-level semantic detection and summarization can be found in the literature for a variety of sports including basketball [19,20], baseball [21,22], Formula-1 [23], tennis [10,24], and American football [25]. Of these genre-specific works, the results reported generally show successful applications but they tend to rely on complex algorithms performing standalone modeling of specific high-level semantic content peculiar to each sport. That is, the content analysis algorithms are embedded within systems and cannot easily be redefined, i.e. recycled for different sports. Little of the prior work has addressed a more generic methodology.

However, while sports video has some common generic characteristics across different sports, it is infeasible to suggest that there exists a unique solution that will operate successfully across all sports. Nonetheless, our approach is pseudo-generic in that it can be applied across multiple sports in the same broad genre. Popular sports genres include athletic track and field, aquatic, ice-sports, and fight sports, although more than half of sports videos produced are some form of field sport game e.g. soccer, rugby, basketball, American football, Gaelic football, Australian rules, field hockey, etc. This makes field sports an ideal genre to target for our pseudo-generic approach.

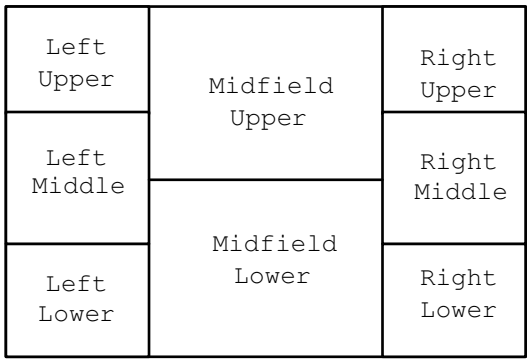


FIGURE 1. The pitch in a field sports game

The characteristics of field sports include:

- Two opposing teams and referee(s);
- A limited number of cameras distributed on three sides of the field of play (the "pitch", shown in Figure 1);
- A limited number of well-defined distinct camera shot classes, such as "Long view", "Medium view", "Tight view" and "Out-of-Field view";
- A limited number of well defined aural and visual characteristics that occur in different types of shots, including opposing teams' players and coaches, referees, captions, a rectangular pitch, field lines, goalposts, a ball, the commentator's voice, the sounds of a whistle and of the crowd cheering;
- Certain techniques of shooting and playback (such as particular kinds of camera motion & change, focus zoom, slow-motion replay, etc.) that are closely related to the semantics of the game.

In this paper, visual, aural and motion features are used to describe the characteristics of field sports video. These are then combined in order to describe higher-level semantic content. For the combination we use Petri-Nets, which describe the relationships between concepts. Petri-Nets are a graphical and mathematical modeling tool [26] which have been used successfully in video meta-data modeling and query processing [27], supporting content-based queries in digital on-demand video services [28], and modeling the synchronization and spatial aspects of multimedia presentations [29]. Using this framework, we can build semantic queries to search within sports videos for specific event types and to generate personalized highlights from that.

As mentioned earlier, we choose to use Petri-Nets because they allow for easy and flexible encoding of the variability that can occur among the sequencing of low-level perception concepts and they achieve this by depending upon accurate detection of these lower-level concepts. We define the visual, aural and motion concepts that we use in the next section of this paper.

3. PERCEPTION CONCEPTS IN SPORTS VIDEO

In generating a summary of a field sports game, a TV sports program editor is interested in selecting specific clips that can help an audience to understand and enjoy the game. For example, rugby highlights might include penalty kicks, trys and drop goals; soccer highlights might include goals and corner kicks, and so on. These elements share similar spatio-temporal behaviour across different perceptual channels, enabling them to be clustered accordingly. In our work, a *Perception Concept (PC)* is defined as the abstraction of video elements extracted from visual, motion and aural perceptual channels which share similar low-level perceptual features. In field sports game videos, there are three types: *Visual Concepts*, *Aural Concepts* and *Motion Concepts*. Our full set of PCs are defined, and the relationships between them are illustrated in Figure 2, and explained in more detail now.

We regard a *Visual Concept (VC)* as including two sub-concepts: visual sequence concepts and visual object concepts. A *Visual Sequence Concept (VSC)* is a sequence of frames captured by one camera in a single continuous action with fixed focus and motion style. In field sports game videos, a VSC can be further classified as one of the following (where the typical relative field view percentages of each may be perceived from the examples provided in Figure 3(a)):

- *Loose View Concept (LVC)*. Based on defined areas of the pitch, the loose view can be further divided into eight kinds of *Pitch Area Concepts(PAC)*: *Left Middle Concept (LMC), Right Middle Concept (RMC), Midfield Upper Concept (MUC), Midfield Lower Concept (MLC), Left Upper Concept (LUC), Left Lower Concept (LLC), Right Upper Concept (RUC)* and *Right Lower Concept (RLC)*;
- *Medium View Concept (MVC)* - i.e. a zoomed-in view, which captures player action at a localized level;
- *Tight View Concept (TVC)* - i.e. a close-up view of a person, in which only the face, shoulders, and torso are visible;
- *Out-of-field View Concept (OVC)*;
- a special kind of visual sequence called a *Slow-Motion-Replay Concept (SMR)*.

A *Visual Object Concept (VOC)* is an image region representing a distinct semantic object in a video frame. In field sports games, important visual objects include

(a) Video Sequence Concepts

(b) Video Objects

FIGURE 3. Visual concepts in field ball games

opposing teams' players and coaches, referees, captions, a rectangular pitch, goalposts, and field lines. In our work, we define and detect three kinds of VOCs (as shown in Figure 3(b)):

- *Caption Object (CO)*;
- *Referee Object (RO)*;
- *Goalpost Object (GO)*

The audio track from a sports video can provide useful indications of a highlight event. For example, when exciting passages of play occur in field games, the crowd and commentators often shout and cheer loudly. An *Aural Concept (AC)* is the abstraction of aural elements which share similar aural features and includes two sub-concepts: aural sequence concepts and aural objects.

An *Aural Sequence Concept (ASC)* is a segmentation of the audio track that belongs to a distinct audio class and lasts a "long time". In our work, we use three kinds of aural sequence concepts, namely *Game-audio Sequence Concepts (GsC)*, *Advertisement-audio Sequence Concepts (AsC)* and *Studio-audio Sequence Concepts (SsC)*. For example, the typical structure of a soccer game video is a first-half, followed by advertisements, studio discussions, advertisements and then the second-half. Usually, the soundtrack of the game is speech with background sound, advertisements are a mixture of speech and music, while in-studio is pure speech.

An *Aural Object (AO)* is a short audio segment that belongs to a distinct audio class and unlike an *Aural Sequence Concept* (ASC), it lasts only for a "short" time. In field ball games, audio cues related to interesting events include a referee's whistle, a loud exclamation from commentators and cheers from the

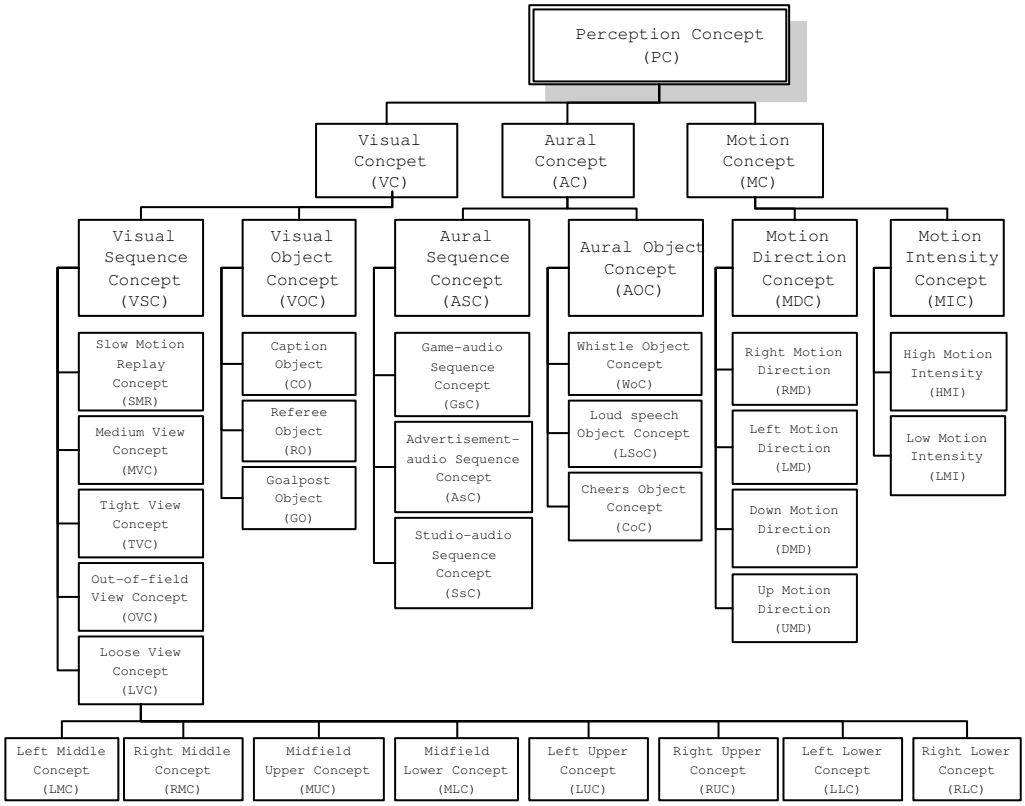


FIGURE 2. Perception Concept Relationships

crowd. We define these three aural objects as *Whistle object Concepts (WoC)*, *Loud Speech object Concepts (LSoC)*, and *Cheers object Concepts (CoC)*.

Finally, *Motion Concepts (MC)* are the abstraction of motion elements corresponding to a continuous camera motion with a distinct direction and motion intensity. We define two types of motion concepts as a *Motion Direction Concept (MDC)* which includes the categorisations *Right Motion Direction (RMD), Left Motion Direction (LMD), Up Motion Direction (UMD)* and *down motion direction (DMD)*, and *Motion Intensity Concept (MIC)* which includes two categories *High Motion Intensity (HMI)* and *Low Motion Intensity (LMI)*.

This collection of nearly 40 perception concepts corresponds to our ontology of recognisable elements for field sports video. In the next section we will examine how these can be detected automatically.

4. DETECTING PERCEPTION CONCEPTS

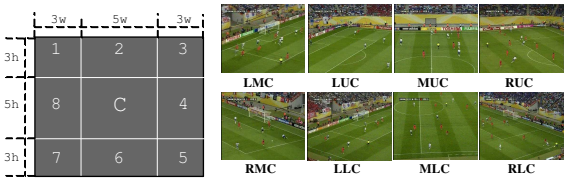
4.1. Detecting Visual Sequences

Like most other work in this area, we initially segment video recordings of sports programmes into shots from which keyframes extracted. This is a well-studied problem with many known solutions which are both fast and effective. Because of the high tempo of live action segments in sports, the broadcast director has little chance to utilize shot transition types other than hard cuts. It has been previously shown that 94% of all shot transitions in field sports are of this nature [1], and so we focus only on detection of hard cuts. We use a mutual information measure between two successive frames calculated separately for each RGB channel, i.e. similar to that reported in [30]. We evaluated the effectiveness of this on a collection of soccer, rugby and basketball matches with an overall duration of 68 mins and 32 seconds and it achieved an overall performance of 95.9% recall and 92.5% precision, which reflects the accuracy of the state of the art [31].

For selecting a keyframe to represent each shot, a number of potential key frames are extracted (i.e. if the shot is 100 frames long, and ten frames are selected, then the 1st, 11th, 21st . . ., 91st frames are potential key frames) and the frame with the color histogram closest to the average histogram over the shot is selected.

4.2. View and Slow-Motion Replay Detection

The topic of camera view detection in sports video has been previously explored, with most emphasis placed on on "field pixel" thresholding as a proposed solution [32, 33, 34]. On this basis, in order to detect the different camera views outlined in Figure 3(a), (LVC, MVC, TVC and OVC), we define regions of the keyframe by dividing the the screen into 3:5:3 proportions in both directions based on the Golden Section spatial composition rule



(a) Golden section spatial composition

(b) Eight PACs in soccer

FIGURE 4. View detection from visual analysis

[35] as shown in Figure 4(a). We then compute the field color pixel ratio R (i.e. the the ratio of the occurrence of dominant field color to that of other colors) for the centre region C (CRR) of each keyframe. If the CRR value is more than a predefined threshold, this sequence is classified as a loose view concept (LVC). The rationale for this is derived from the fact that editors like to frame important objects in a medium or tight view in the C region of the screen. In our previous work, a color-based spatial model was proposed in [1] for TVC detection, where skin color ratio, shirt color ratio and background color ratio are computed, and a formula for image close-up confidence is defined. Finally, we calculate the field colored pixel ratio to classify out-of-field (OVC) views and medium views (MVC).

As outlined in section 3, one *Loose View Concept (LVC)* often contains several different *Pitch Area Concepts (PAC)* due to a panning camera capturing different parts of the field of play in one shot. We can abstract the important characteristics of different PACs that represent the spatial relationships between pitch, crowd and screen. For example, in a *Left Upper Concept (LUC)* the crowd will always appear in both the left and upper parts of the screen. Different areas of the pitch are shown in Figure 4(b). Initially, each LVC is segmented into half-second clips and the middle frame of the clip is selected as the representative frame. We then calculate the field color pixel ratio, R, in regions No.1 to No.8 as defined in Fig. 4(a). Each representative frame can thus be expressed by a representative vector with 8 components. A K-means clustering algorithm, with 8 clusters, is used to classify the representative vectors. Finally, in each LVC, consecutive clips of the same class, are merged.

For detecting *Slow-Motion Replay Concepts (SMR)*, we use a zero crossing measure which is aligned with the technique outlined in [15].

4.3. Detecting Visual Objects

4.3.1. Caption Detection

Caption Objects (CO) often occur after important events in field sport games, such as a scoreboard appearing after a score, or player information caption after a yellow card event. Captions are always located in a fixed part of the screen, and there will be some corners located in the caption region, and these corners will have the same locations in consecutive frames in which a caption occurs. Figure 5(a) shows examples of corners with fixed locations (red tags) in semi-consecutive frames (one frame per 10 frames) in soccer video. Based on our observations, captions tend to last more than 3 seconds. We use a simple and effective method to detect *Caption Objects* and avoid calculation in consecutive frames. At first, we select the first frame in every 10-frames and use the Harris algorithm [36] to detect re-occuring corners. The number of corners in each frame is counted and if greater than a predefined threshold we determine the occurrence of a caption object.

4.3.2. Referee Detection

In field game videos, the occurrence of a *Referee Object (RO)* is an important cue indicating interesting events. For example, after a foul event the broadcasting editor often shows the referee(s) in a medium or tight view. Referees wear distinguishable color uniforms from those of the two teams on the field and we exploit this towards RO detection. Firstly, we construct color templates of the referee based on training data using RGB color histograms. A given test frame is divided into non-overlapping blocks and each block's histogram is compared to the template. Incorrect matches are filtered based on checking the block's 8-neighbourhood for non-RO blocks.

4.3.3. Goalpost Detection

Goalpost objects (GO) often occur in scoring attempts in field sports and thus the occurrence of a GO is another important cue indicating these types of events. A GO in field games is a pair of white vertical parallels of fixed width. Using the luminance channel of the frame, vertical edges can be detected using Sobel's algorithm. After noise removal, a Hough transform is used to detect vertical parallels which indicate the detected GOs. An example of GO detection is shown in Figure 5(b).

4.4. Detecting Aural Concepts

In order to detect the three kinds of aural *sequence* concepts defined in Figure 3(a) (GsC, AsC and SsC), four audio features and an SVM are used. The features used are silence ratio, high zero-crossing rate ratio, low frequency energy ratio and spectrum flux as reported in our previous work [37][38][39].

To detect the various audio *objects* defined in Figure 3(a) (WOC, LSoC, CoC), we calculate the audio frequency energy (FE), defined as the total spectrum power of a frame and the frequency centroid (FC) of the spectrum in a frame. We use a dual-threshold method

(a) Caption object (CO) corners with fixed locations on-screen

(b) A example of goalpost object (GO) detection

FIGURE 5. Visual object detection

to detect WoCs. A frame is determined as a possible start point (PSP) if its FE and FC values are greater than the pre-defined thresholds. Then the average values of FE and FC in 70 consecutive frames following the PSP are calculated. If the average values of FE and FC are more than the pre-defined thresholds, a WoC is detected. LSoC is a speech signal with high energy. CoC always has low frequency and high energy. The mean value of FE and FC in one clip are calculated and three features are extracted namely high zero-crossing rate ratio, low frequency energy ratio and spectrum flux. Two SVM classifiers are trained and used to detect LSoC and CoC respectively.

4.5. Detecting Motion Concepts

In general, when a fast break or counterattack and a long pass occur in a field sports game, the broadcasting editor uses a loose view to capture the whole situation on the pitch in order to contextualize the action taking place. On the other hand, a medium view is typically used to capture motion intensive action such a tackle. Hence in this work, motion direction concepts are only detected in *Loose View Concepts (LVCs)* and motion intensity concepts are only detected in *Medium View Concepts (MVCs)*.

To detect the different motion direction concepts (up, down right and left motion directions, represented as UMD, DMD, RMD and LMD) defined in Figure 3(a), we use MPEG B-frame motion vectors. Before use, the motion vectors are median-filtered to eliminate rogue vectors. (Note, the fidelity of MPEG motion vectors to true video motion is questionable, due to the fact that they simply represent the results of mathematical pixel-block matching techniques between successive images. Hence the need for reliable filtering to eliminate the presence of outliers.) Once filtered, we categorize each (non-zero) motion vector as either vertical+ (i.e. upward motion), vertical- (i.e. downward motion), horizontal+ (i.e. rightward motion), or horizontal- (i.e. leftward motion). The most occupied category then indicates the overall camera motion direction.

Motion intensity is measured for each detected medium view sequence as follows. It is well known that the standard deviation of motion vector magnitudes offers optimum fidelity to the true motion of a video sequence [40]. On this basis, we calculate the mean and standard deviation of each P-frame, then the mean of these standard deviations. If the mean value is more than a predefined threshold, *High Motion Intensity (HMI)* is detected, otherwise *Low Motion Intensity (LMI)* is detected.

Before we move on an introduce our PCN-PN approach to modelling sports events, it must be acknowledged that automatic detection of mid-level and high-level features in video, and in still images, is not an exact science and even the best detection techniques will have errors in their output. Since 2001 the TRECVid benchmarking activity has been field testing the effectiveness of various feature detection methods from dozens of research groups, on hundreds of hours of video each year. Over the last 7 years we have seen a marked and significant improvement in the effectiveness of the best of these techniques [41], and for some of the low-level features such as camera motion and object motion intensity and direction, detection

of goalposts, captions, and some aural sequences, the performance of state-of-the-art detection techniques is certainly usable and in some cases almost perfect. While not necessarily novel or original, the techniques we use to detect the variety of perception concepts we described in this section of the paper represent the state-of-the-art in terms of techniques and performance figures and provide us with output results which are stable and reliable enough for us to build PCN-PNs, which we describe in the next section.

5. DEFINING AND USING PCN-PN MODELS

As summarized in [42], a Petri-Net is a graphical modeling tool consisting of places, transitions, and arcs for connecting them. Input arcs connect places with transitions, while output arcs start at a transition and end at a place. Places can contain tokens. The current state of the modeled system (the marking) is given by the number (and type if the tokens are distinguishable) of tokens in each place. Transitions are active components which model possible activities (when the transition fires), thus changing the state of the system (the marking of the Petri-Net). Transitions are only allowed to fire if they are enabled, which means that all the preconditions for the activity must be fulfilled (i.e. there are enough tokens available in the input places). When the transition fires, it removes tokens from its input places and adds to some or all of its output places.

Based on the definition of *Perception Concepts (PCs)* in Section 3, a field sports video is composed of a set of PCs and their relationships, and "interesting" semantic content such as events can be represented by a fixed set of PCs and their relationships. For example, when a goal is scored a consistent and predictable combination of shots is observed: from *Loose View Concept (LVC)* to a *Tight View Concept (TVC)*, an *Out-of-field View Concept (OVC)*, a *Slow Motion Replay (SMR)*, and another player TVC, finally returning to a LVC, whilst at the same time, a *Cheering Object Concept (CoC)* and a *Caption Object (CO)* are often also detected. In our work, each PC is represented by a token and the relationships between PCs are modeled by Petri-Nets. The field sports video is preprocessed to detect the set of low-level PCs within it. The initial input places in the Petri-Net represent all the types of PCs that can contribute to the detection of a high-level event and each token initially assigned to these places represents an occurrence of a low-level PC detected within the video. After the Petri-Net has completed execution, i.e. when no transitions can fire, the tokens assigned to the final output place represent the occurrences of the high-level semantic event. Each token will define the time boundary, start time and end time, of the high-level semantic event. Based on this idea, we define a Perception Concept Net based on the Petri-Net model (PCN-PN) as follows

Definition 5.1 A token is an object that represents a low-level *Perception Concept (PC)* that starts at time t_{start} and ends at time t_{end}.

$$Tok = \{tok_i(t_{start}, t_{end}) \mid i \geq 0, t_{start} \geq 0, t_{end} \geq t_{start}\}$$
$$start(tok(t_{start}, t_{end})) = t_{start}$$
$$end(tok(t_{start}, t_{end})) = t_{end}$$

Definition 5.2 A PCN-PN is a 6-tuple

$$C_{PCN-PN} = \{T, P, A, Op, Du, M\},$$

where:

$T = \{t_1, t_2, \cdots, t_n\}$ is a finite set of transition with $n \geq 0$;

$P = \{p_1, p_2, \cdots, p_m\}$ is a finite set of places with $m \geq 0$, and $P \cap T = \emptyset$. Especially, *Null* place is defined to describe a short circuit of two transitions because two transitions can not be connected directly, which is expressed by a blank rectangle graphically;

$A : (P \times T) \cup (T \times P) \longrightarrow I, I = \{1, 2, \cdots\}$ is a set of arcs from places to transitions and from transitions to places;

$Op : T \longrightarrow Operator$ where $Operator = \{and, or, following^i, before^i, synchronization\}$ is a mapping from transition to its operators set and $i = \{1, 2, \cdots\}$;

$Du : P \longrightarrow (Operate, Time)$ is a mapping from place to its duration, $Operate \in \{<, =, >\}, Time \in R$;

$M : P \longrightarrow Tok$ is a mapping from place to its marking, i.e. the set of tokens assigned to each place.

Each transition has an operator mapped to it. The operator defines the conditions under which the transition is enabled and the conditions under which it fires.

Definition 5.3 Given a transition t, the set of input places, $\bullet t$, and successor places, $t\bullet$ are

$$\bullet t = \{p \mid p \in P \wedge (p, t) \in A\}$$
$$t\bullet = \{p \mid p \in P \wedge (t, p) \in A\}$$

Definition 5.4 ***and*** **Transition** A transition t mapped to an *and* operator is enabled and fires when all input places have one or more tokens. When the transition fires it removes a token from each input place and adds a token to each output place. The t_{start} value of the new tokens is equal to the minimum value of the t_{start} values removed from the input places. The t_{end} value of the new tokens is equal to the maximum value of the t_{end} values removed from the input places.

enabled: $\forall p_i \in \bullet t, M(p_i) \neq \emptyset$
fires: *true*
result: $\forall p_i \in \bullet t, tok_j \in M(p_i), p_k \in t \bullet .$
$M(p_i) \rightarrow M(p_i) - \{tok_j(t_s, t_e)\}$
$t_{min} = min(\{t_s\}), t_{max} = max(\{t_e\})$
$M(p-k) \rightarrow M(p_k) \cup \{tok(t_{min}, t_{max})\}$

Definition 5.5 *or* Transition A transition t mapped to an *or* operator is enabled and fires when any of the input places has one or more tokens. When the transition fires it removes a token from each non-empty input place adds a token to each output place. The t_{start} value of the new tokens is equal to the minimum value of the t_{start} values removed from the input places. The t_{end} value of the new tokens is equal to the maximum value of the t_{end} values removed from the input places.

enabled: $\exists p_i \in \bullet t, M(p_i) \neq \emptyset$
fires: *true*
result: $\forall p_i \in \bullet t, tok_j \in M(p_i), p_k \in t \bullet .$
$M(p_i) \rightarrow M(p_i) - \{tok_j(t_s, t_e)\}$
$t_{min} = min(\{t_s\}), t_{max} = max(\{t_e\})$
$M(p-k) \rightarrow M(p_k) \cup \{tok(t_{min}, t_{max})\}$

Definition 5.6 *following*^{i} Transition A transition t mapped to a $following^i$ operator is enabled when all input places have one or more tokens. Unlike the *and* and *or* transitions, the *following*^{i} transition is an asymmetric binary transition with 2 input places. Given input places p_A and p_B then *following*^{1} is defined by,

enabled: $\forall p_i \in \bullet t, M(p_i) \neq \emptyset$
fires: $\exists i, j \forall k \mid tok_i \in p_A \wedge tok_j \in p_B \wedge tok_k \in \bullet t$
$\wedge start(tok_j) \geq end(tok_i)$
$\wedge \neg(end(tok_i) \leq start(tok_k)$
$\wedge end(tok_k) \leq start(tok_j))$
result: $p_A \rightarrow p_A - \{tok_i\}, p_B \rightarrow p_B - \{tok_j\}$
$\forall p_l \in t \bullet .$
$p_l \rightarrow p_l \cup \{tok(start(tok_i), end(tok_j))\}$

Graphically the place p_A is above the place p_B.

$following^i$ is fired when there are $(i-1)$ sequential tokens in $p_A \cup p_B$ between a token in p_A and a token in p_B. The enabling condition and effects are the same as $following^1$. $following^*$ is a generalisation of $following^i$ and will fire if there is any sequence of sequential tokens between a token in p_A and a token in p_B.

Definition 5.7 *before*^{i} Transition A transition t mapped to a $before^i$ operator is enabled when all input places have one or more tokens. The *before*^{i} transition is an asymmetric binary transition with 2 input places. Given input places p_A and p_B then *before*^{1} is defined by,

enabled: $\forall p_i \in \bullet t, M(p_i) \neq \emptyset$
fires: $\exists i, j \forall k \mid tok_i \in p_A \wedge tok_j \in p_B \wedge tok_k \in \bullet t$
$\wedge end(tok_j) \leq start(tok_i)$
$\wedge \neg(end(tok_j) \leq start(tok_k)$
$\wedge end(tok_k) \leq start(tok_i))$
result: $p_A \rightarrow p_A - \{tok_i\}, p_B \rightarrow p_B - \{tok_j\}$
$\forall p_l \in t \bullet .$
$p_l \rightarrow p_l \cup \{tok(start(tok_j), end(tok_i))\}$

Graphically the place p_A is above the place p_B.

$before^i$ is fired when there are $(i-1)$ sequential tokens in $p_A \cup p_B$ between a token in p_B and a token in p_A. The enabling condition and effects are the same as $before^1$.

Definition 5.8 *synchronization* Transition A transition t mapped to a *synchronization* operator is enabled when all input places have one or more tokens. The *synchronization* transition is an asymmetric binary transition with 2 input places. Given input places p_A and p_B then *synchronization* is defined by,

enabled: $\forall p_i \in \bullet t, M(p_i) \neq \emptyset$
fires: $\exists i, j \mid tok_i \in p_A \wedge tok_j \in p_B$
$\wedge start(tok_i) \leq start(tok_j)$
$\wedge end(tok_i) \geq end(tok_j)$
result: $p_A \rightarrow p_A - \{tok_i\}, p_B \rightarrow p_B - \{tok_j\}$
$\forall p_l \in t \bullet .$
$p_l \rightarrow p_l \cup \{tok(start(tok_i), end(tok_i))\}$

Graphically the place p_A is above the place p_B.

Typically the token in the input place p_A represents a Video Sequence Concept (VCS) and the function of the *synchronization* transition is to ensure that a Visual Object Concept (VOC), Aural Concept (AC) or Motion Concept (MC) occur within the VSC.

In previous work, semantic content analysis and event detection in video mainly focused on detecting general highlights or general events. This works effectively to a certain level but can not satisfy users' needs for personalized *semantic* content access. Sometimes users want to see specific event types and generic highlight detection cannot facilitate this. However, semantic content in field sports game videos can be flexibly described using the PCN-PN model. That is, by using Petri-Nets, perception concepts (PCs) and their relationships can be combined at different complexity levels to construct a PC net that describes different kinds of complex semantic content. This makes it possible that users can use network management tools to design their own semantic content descriptions based on the PCN-PN model and generate their own personalized highlights of videos

Note that it is intuitive to describe and search semantic content using combinations of perception concepts. Although the definition of a PC is based on the consistency of low-level features, it will contain

simple and easily understood semantics. For example, a *Tight View Concept (TVC)* is often used to highlight important people in a game; *Slow Motion Replay (SMR)* concepts often record and replay scoring attempts, etc. Highlights based on low-level semantics, such as "all in-play shots in the right field in a game", "all SMR segments in a game" or "all shots of the referee close-up", can be processed directly based on PCs.

The combination of some single PCs can also be used to represent mid-level semantics. For example, a "long pass" from left upper field to right upper field can be decomposed into three different PCs, a *Left Upper Concept*, a *Midfield Upper Concept* and a *Right Upper Concept* (LUC + MUC + RUC). As another example, a "referee close-up" can be decomposed into a *Tight View Concept* and a *Referee Object* (TVC + RO). The symbol "+" here represents some kind of relationship between PCs. According to the definitions of the PCN-PN model, the relationships between PCs are represented by the element *Op* defined in Definition 5.2. The "Long pass" and "Referee close-up" semantics described above can be described as in Figure 6.

High-level semantics in field sports videos can be represented by combinations of PCs. The PCN-PN description of a scored goal in soccer is shown in Figure 7(a). This shows a general description of scored goal event. If a particular type of scored goal event, such as a scored goal derived from a right side attack is required, then we can flexibly define a new PCN-PN. This simply requires modifying the left-lower part of the PCN-PN model as shown in Figure 7(b). The PCN-PN description of a "Foul" in soccer is shown in Figure 8. The PCN-PN description of a fast break in basketball from the left court to the right court is shown in Figure 9. The new PACs in the basketball game used in Figure 9, LHC, MHC and RHC will be described in the next section. The PCN-PN description of a try in rugby is shown in Figure 10.

6. EXPERIMENTS AND EVALUATION

The defined Petri Net patterns are stored in a directed graph. While there are probably several candidates for representing the low-level patterns that make up a sporting event such as regular expressions for example, we choose to use Petri Net representations because of the flexibility they offer in terms of re-use of components. By this we mean that out approach is more generic than, say, using regular expressions because the re-use of low-level perception concepts is easier when using Petri Nets as we have found when re-using low-level concepts in moving from soccer to rugby and to other field sports. Before comparing the Petri Net patterns against a recorded match, all Perception Concepts in the match are detected and stored. For a given PCN-PN of a semantic event, we search for the start PC and followed by the end PC in the detected PCNs, and a pair of start-end segments is regarded a candidate event. We then search the whole path from start PC to end PC in the stored graph model of the given PCN-PN. If the search returns true, the candidate is decided as a true event.

TABLE 2. SMR detection results

Correct	False	Miss	Recall	Precision
157	23	35	82.8%	87.2%

The time taken to do this computation is composed of two parts: one is the detection of PCs, the second is searching for given PCN-PNs. Taking soccer as an example, the time taken to detect PCs in a match, including visual, aural and motion, is about 30 minutes, which is a few times faster than real time on a standard desktop computer.. The time taken to search for a given PCN-PN will of course be dependent on the complexity of the PCN-PN and the candidate segments as well as whether the video can be searched within the compressed domain or needs to be decompressed. The search for simple PCN-PNs, such as low and middle level semantics, can be completed in just a few seconds. For the most complex PCN-PNs, such as goal semantics, it can take several minutes.

In order to demonstrate the validity and usefulness of our approach we conducted a set of experiments using a number of soccer, rugby and basketball matches encoded as MPEG-1. The sports videos are from a range of broadcasters, and consist of over 15-hours of footage. Table 1 shows the details of the experimental data set.

6.1. Results for Perception Concept (PC) Detection

6.1.1. VSC detection results

Visual Sequence Concept (VSC) detection includes three parts, SMR (slow motion replay) detection, LVC-MVC-TVC-OVC (loose, medium, tight and out-of-field view) detection and *Pitch Area Concept (PAC)* detection. SMRs are detected at the end of the first half of each soccer and rugby game and at the end of the first quarter in each basketball game and the results are shown in Table 2 achieving 82.8% recall and 87.2% precision. These results are worse than the reported results, 100% recall rate without a precision rate, reported in [15] though that work has been done on a different dataset to ours.

Eight video sequences segmented from eight test videos were selected as the ground truth set for evaluating LVC, MVC, TVC and OVC detection, the details of which are as follows: S1(20m:10s), S2(17m:29s), S3(22m:42s), B1(22m:34s), B2(27m:43s), B3(24m:30s), R1(25m:48s) and R2(21m:25s). Because the characteristics of a basketball court are very different from that of a pitch in soccer or rugby,

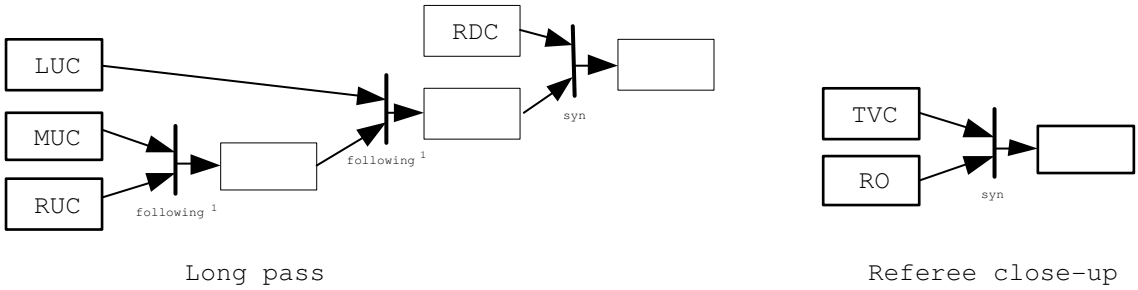


FIGURE 6. Examples of mid-level semantic descriptions using a PCN-PN in soccer

TABLE 1. Video data for experiments

ID	Name	Broadcaster	Duration
S1	2006 World Cup FRA vs KOR	BBC Sports	1h:35m:36s
S2	2006 World Cup POR vs NED	ITV Sports	1h:39m:44s
S3	2006 World Cup GER vs SWE	BBC	1h:35m:25s
B1	NBA06-07 Rocket vs Jazz	ESPN Live	1h:53m:26s
B2	NBA06-07 Rocket vs Spurs	CCTV5	1h:50m:14s
B3	NBA06-07 Rocket vs Magic	Fox Sports	2h:10m:46s
R1	Rugby World Cup 07 NZL vs ITA	SETANTA	1h:31m:24s
R2	Rugby World Cup 07 AUS vs JPN	SETANTA	1h:29m:16s

(a) *(b)* (c) *(d)*

FIGURE 11. Examples of missing and false in visual sequence concept (VSC) detection

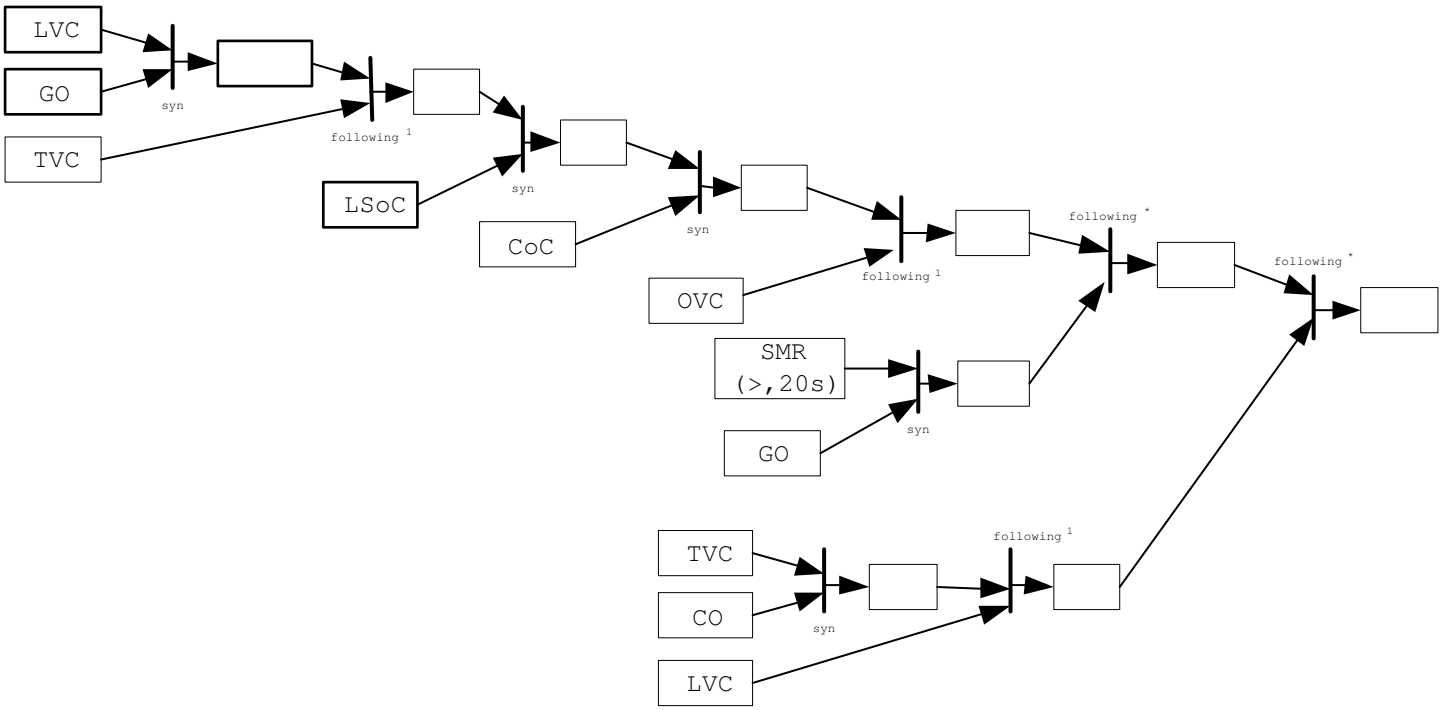
TABLE 3. LVC, MVC, TVC and OVC detection results

	VSCs	Correct	False	Miss	Recall	Precision
S/R	LVC	315	11	0	100%	96.6%
	MVC	213	16	15	93.4%	93.0%
	TVC	156	0	21	88.1%	100%
	OVC	15	9	0	100%	62.5%
	ALL	699	36	36	93.8%	
B	LVC	158	0	42	79.0%	100%
	MVC	23	45	4	85.2%	33.8%
	TVC	71	0	10	87.7%	100%
	OVC	9	11	0	100%	45%
	ALL	261	56	56	82.3%	

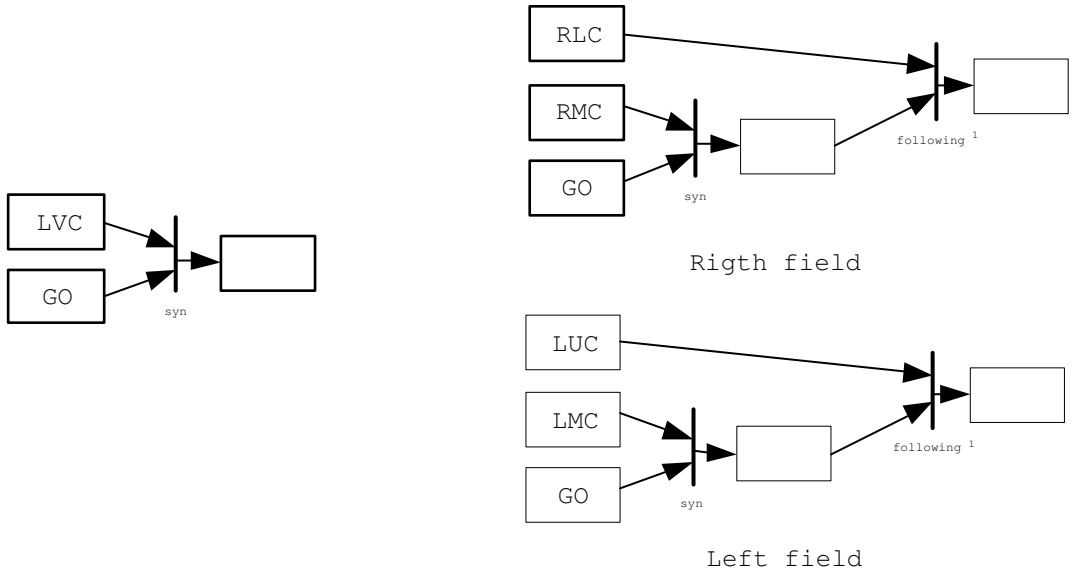
B denotes basketball and S/R denotes soccer and rugby.

our algorithms for VSC detection are evaluated on basketball videos and soccer/rugby videos separately. The results are shown in Table 3 where we find that much better recall and precision rates are achieved in S/R experiments than in B experiments. The main reason for this is because of color characteristics in basketball videos which are different from other field sports games. Different parts of a basketball court often have different colors and our algorithms designed based on the color characteristics do not work well in such cases. Many misses in TVC detection are due to capturing a tight view from back or side face, as shown in Figure 11 (a) and (b). The missed TVCs are incorrectly determined as OVC because of low field colored pixel ratios or as MVC incorrectly because of high field colored pixel ratios. Some MVCs have non-field backgrounds as shown in Fig.11 (c) and (d) and are detected as OVC incorrectly. The recall and precision rates in S/R experiments are more satisfactory.

The *Pitch Area Concepts (PACs)* are detected in the detected *Loose View Concepts (LVCs)*. Because the pitch in a basketball game is much smaller than that in soccer and rugby games, the definitions of PACs in a basketball game are different from the eight PAC definitions in section 3. According to the characteristics of a basketball pitch, we define three kinds of PACs in basketball game videos as a *Left Half Concept (LHC)*, *Right Half Concept (RHC)* and *Middle Field Concept (MFC)*. We use the same PAC detection approach described earlier for basketball video PAC detection. The results are shown in Table 4.



(a) General semantic description of a "scored goal" in soccer



(b) A "right side attack goal" in soccer

FIGURE 7. Semantic descriptions of a "scored goal" in soccer

6.1.2. VOC detection results

In basketball and rugby games *Visual Object Concepts (VOCs)* such as the *Referee Object* (RO) does not always occur in a *Tight View Concept* (TVC), and so ROs are only detected in soccer videos. Goalpost objects (GOs) are only detected in soccer and rugby videos because there are no GOs in basketball. The ground truth set for VOC detection was obtained from the test video sequences used in VSC detection and the results of VOC detection are shown in Table 5. Our RO detection algorithm achieves good performance. When capturing local scenes in medium views (MVC),

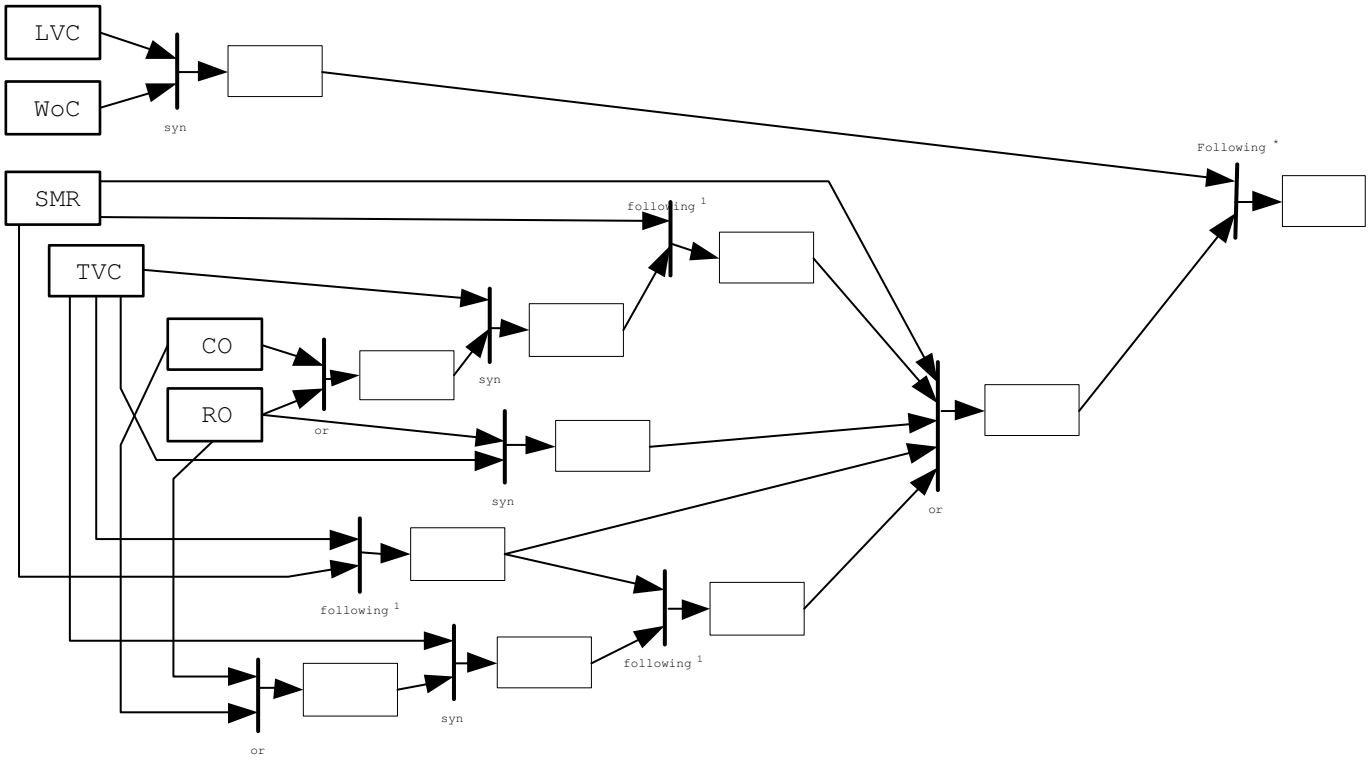


FIGURE 8. Semantic description of a soccer "foul"

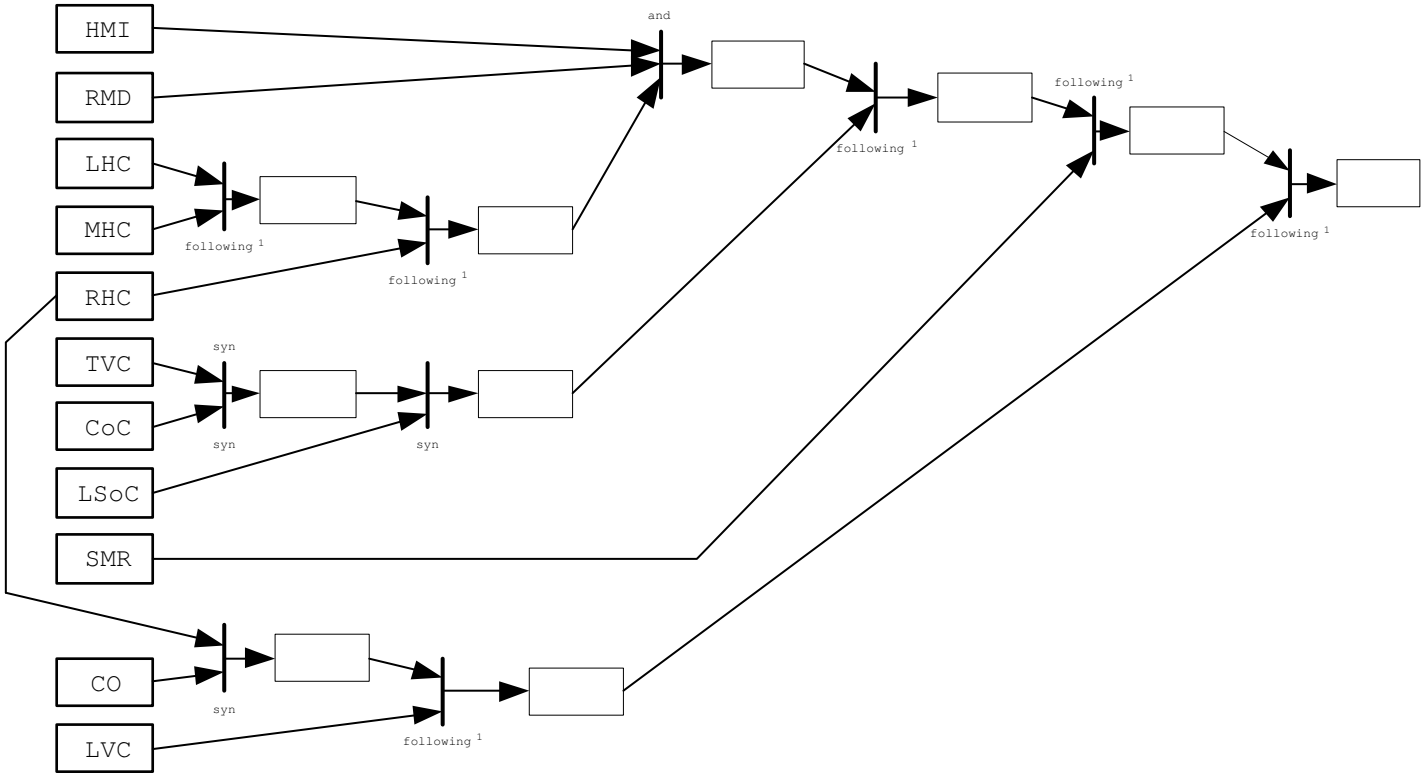


FIGURE 9. Semantic description of a "fast break" in basketball

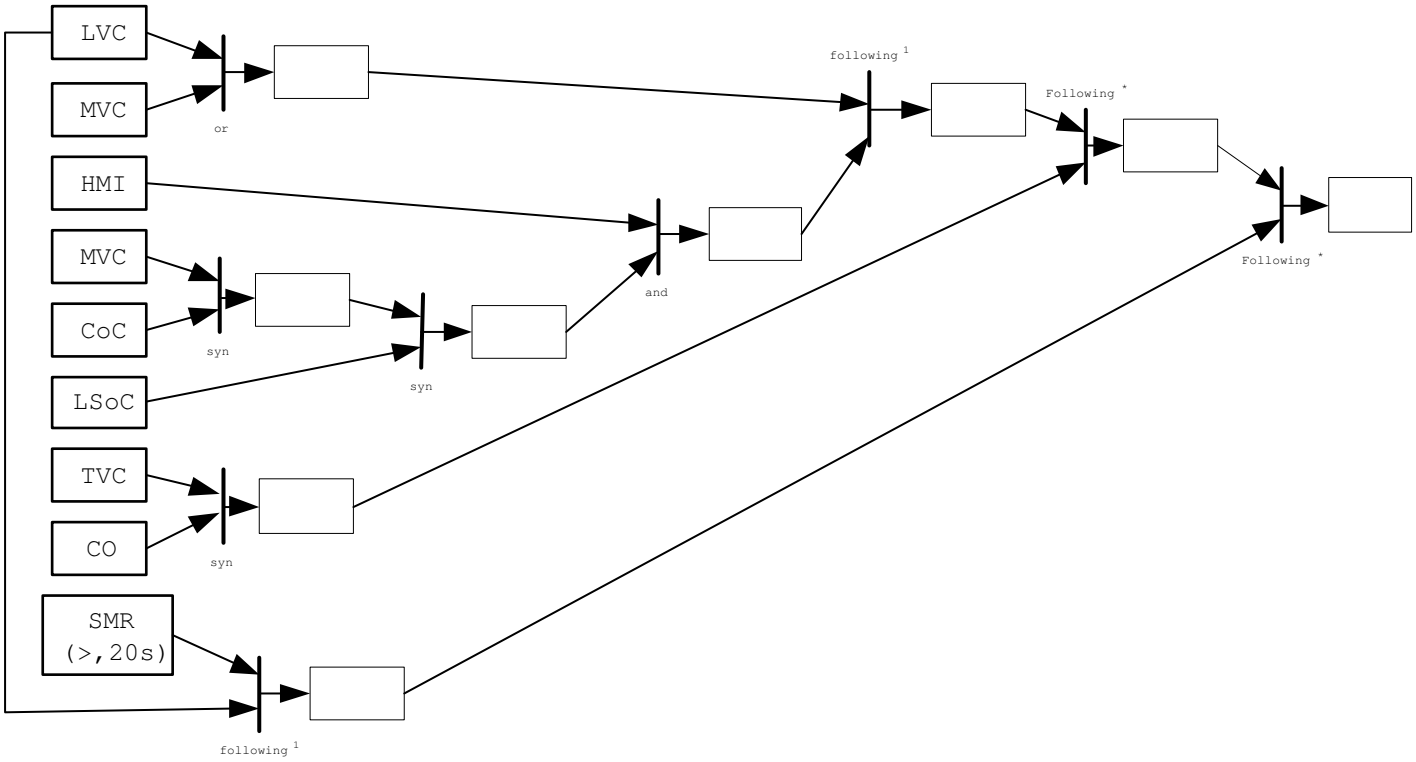


FIGURE 10. Semantic description of a rugby "try"

TABLE 4. PAC detection results

PACs	Correct	False	Accuracy
8-PACs	739	93	88.8%
B-PACs	324	67	82.9%

8-PACs denotes the PACs in soccer and rugby videos; B-PACs denotes the PACs in basketball video

TABLE 5. VOC detection results

VOCs	Correct	False	Miss	Recall	Precision
RO	29	0	0	100%	100%
CO	143	11	0	100%	92.9%
GO	65	0	8	89%	100%

editors sometimes keep the camera static. In such cases, many static corners in advertisement boards or field lines are detected and a *Caption Object (CO)* is determined incorrectly. Some GOs were missed because the background color is similar to the GO's color.

6.1.3. AC detection results

We have discussed audio classification and segmentation for sports video structure extraction in our previous work [39], which is essentially the same problem as *Aural Sequence Concept (ASC)* detection. So in this paper, when we consider *Aural Concepts (ACs)* we do not report further experiments on ASC detection and we focus on aural *Object Concept (AOC)* detection experiments only. Firstly, audio tracks are extracted from each match video. The ground truth set for AOC detection is obtained from S1, B1 and R1. The results for *Whistle Object Concept (WoC)* detection are shown in Table 6. False results are due to the crowd making whistles and other sounds whose spectrum is similar the WoC in play. Missing WoCs are mainly covered up by background noise with high energy and frequency.

TABLE 6. WoC detection results

AOC	Correct	False	Miss	Recall	Precision
WoC	377	98	110	77.4%	79.4%

For loud speech object (LSoC) and *Cheering Object Concept (CoC)* detections, two SVMs are trained on one game and tested on others. The result is measured by classification accuracy defined as the number of correctly classified clips over the total number of clips. Training and testing accuracies for LSoC and CoC detections are shown in Table 7. The average classification accuracy (avg-cla) of each program as testing data is computed as the mean of elements of the current row; similarly, the average generalization accuracy (avg-gen) is computed

TABLE 7. LSoC and CoC detection results

Testing data		Training data			avg-cla
		S1	B1	R1	
S1	LSoC	86.6%	84.9%	85.7%	85.7%
	CoC	87.2%	85.8%	86.3%	86.4%
B1	LSoC	85.2%	87.4%	84.7%	85.8%
	CoC	87.5%	86.9%	83.2%	85.9%
R1	LSoC	86.7%	84.1%	86.4%	85.7%
	CoC	84.3%	83.8%	85.2%	84.4%
avg-gen		86.3%	85.5%	85.3%	85.7%

TABLE 8. MC detection results

	MCs	Correct	False	Accuracy
MIC	HMI	87	13	87.0%
	LMI	43	7	86.0%
MDC		923	137	87.1%

for the program as training data and the overall average classification/generalization accuracy over the entire dataset is in the lower right corner. From Table 7, we see that LSoC and CoC detection perform satisfactorily.

6.1.4. MC detection results

The final group of *Perception Concepts (PCs)* whose detection we evaluated were *Visual Motion Concepts (MCs)*, divided into *Motion Direction* and *Motion Intensity Concepts* (MDC and MIC). MIC consisted of *High Motion Intensity (HMI)* and *Low Motion Intensity (LMI)*. The ground truth set we used for the evaluation was obtained from correctly detected *Medium View Concepts (MVCs)*. There is no standard for annotating ground truth of HMI and LMI manually as it is totally subjective, so in our work, we selected 100 MVCs in an "in-play" state and 50 MVCs in "out-of-play" state, and we associated HMI with in-play MVCs and LMI with out-of-play MVCs. The *Motion Direction Concepts (MDCs)* we used as ground truth were detected from the correctly detected *Loose View Concepts (LVCs)*. In each LVC, we manually measured the actual camera motion and compared it against the estimated motion. The results of MC detection are shown in Table 8. False detections of HMI are due to static MVCs which cause incorrect *Caption Object (CO)* detections. Always, in out-of-play scenes there are no strong camera motions, but in some cases, such as players' celebrations or when warming-up, HMIs will occur. False results in MDC detection are mainly due to the camera being shaky because the broadcaster has chosen to use handheld cameras.

6.2. High-level Semantic Event Detection and Highlight Generation

In this section, we focus on demonstrating the validity of the PCN-PN approach for event detection and highlight generation. Note, whilst the experiments on PC detections are useful for demonstrating the feasibility of fully automated processing at concept level, given that errors in automatic PC detection will affect the performance of PCN-PN approach (which would ultimately prevent us from assessing its performance as an event detection paradigm), we first eliminate these errors manually to proceed with the validity assessment of PCN-PN. The designs of PCN-PN models for different high-level semantic events are based on the designers' knowledge of the sports domain and characteristics of sports video broadcasting and editing. In our work, PCN-PN models are developed based on observations from sports videos and the semantic event detections are tested on the experimental video data shown in Table 1. The high-level semantics described by PCN-PN models include *Scored Goal (SG)* and *Yellow* (or *Red*) *Cards* (YC) in soccer games, *Fast Break (FB)* and *Fouls (FL)* in basketball, and *Try (TY)* in rugby games. Table 9 shows "Precision" and "Recall" for detection of these 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.

From Table 9, we can see that the approach achieves good precision and perfect recall rates on *Scored Goal*, *Cards* and *Trys* (SG, YC, TY). The four false detections in yellow card detections are due to 1 debated foul and 3 offside decisions which have a similar model to the yellow card event. Five rugby attacks which did not yield trys are determined incorrectly as try events, though users may be interested in these near-scoring events anyway. Poorer recall and precision rates in *Fast Break* and *Fouls* in basketball (FB and FL) detections are shown in Table 9. Missing FB events are due to the absence of audio cues in some FB events. Successful defense against *Fast Breaks* and non-scoring *Fast Break* events in basketball have a similar model to FB events and are determined as such. The number of non-foul events in the foul detection is proportional to the frequency of the breaks in the game that are due to out of bound, three-second violation, traveling, illegal defense and so on. These non-foul events have a similar model to foul events and cause false FL detections. Missing FLs are mainly due to the absence of TVC and CO perception concepts that are important elements in the PCN-PN description of FLs.

Using detected *Perception Concepts (PCs)* and high-level semantic content, we can generate different types of highlights as outlined below.

Using only one kind of *Perception Concept (PC)* alone, we can construct highlights for a specific event such as all SMRs, which is used as a type

TABLE 9. Results for 5 high-level semantic event detections

Semantic	Actual Num	True Num	False Num	Recall	Precision
SG	5	5	0	100%	100%
YC	23	23	4	85.2%	100%
FB	28	20	12	62.5%	71.4%
FL	69	54	31	63.5%	78.3%
TY	26	26	5	83.9%	100%

of summarization in [4]. Other examples include a highlight being composed of all close-up shots with referees or a highlight composed of all in-play medium view shots which can represent physical play scenes.

More often, however, high-level semantic content in sports video can also mean events and scenes. Based on the high-level semantic content detection described earlier, we can generate highlights composed of the same or different events and scenes, such as all scored goal highlights in soccer, all try event highlights in rugby and so on.

Finally, our framework can allow the design of semantic detectors for many kinds of events, each based on PCN-PNs to search, perhaps interactively, within sports videos. Thus users can have personalized highlights containing infrequently occurring specific semantic content that they want to include, such as all left side attack highlights.

7. CONCLUSION

In this paper, a novel framework for semantic description of field sports video based on *Perception Concepts (PC)* and Petri-Nets has been outlined. Elements located in the visual and aural channels in sports videos which share similar perceptual features are defined as PCs. Three types of PCs, visual, aural and motion, are defined according to the characteristics of sports videos. The algorithms for PC detection are also described, and a high-level semantic description model which integrates PCs into a Petri-Net model is formally defined and called "PCN-PN". As we stated earlier, successful alternative approaches to detection of high-level semantic events in field sports video are based on algorithms which perform standalone modeling of specific high-level semantic content which is peculiar to each sport. Our approach is more generic in that the low-level perception concepts that we detect are re-used for detection of different semantic events both within a given sport, and across different field sports as well.

Semantic detection of video highlights using the PCN-PN model has also been designed and presented in this paper and a data set of more than 15 hours of video, which was obtained from three types of field games and multiple broadcast sources, was used for evaluating the proposed PC detection algorithms and the validity of the PCN-PN model. Compared to the manually annotated ground truth sets, it has been shown that satisfactory results from both PC detection and high level semantics identification, are achieved.

Future work will focus on four main aspects. Firstly, further improvements for PC detection algorithms are required that can improve and ensure good PCN-PN performance in cases of automatic processing. Secondly, while we have demonstrated the feasibility of Petri-Net modeling on a small set events, the scalability of the proposed approach should be further investigated. Thirdly, interface design which is very important to enable rapid development of new event detections is desirable. In addition, a detailed investigation, which compares the approaches herein with other high-level semantic retrieval methods would be beneficial.

Finally, while the emphasis of this work is on building specialised Petri-Nets for automatic event detection, an alternative paradigm may be to facilitate direct end-user querying using Petri-Nets which have been built by the users themselves. However, it should be noted that the construction of customized Petri-Nets is a non-trivial task, and pursuers of the development of such a scheme may want to refer themselves to some related work in Petri-Net construction such as [42, 43].

ACKNOWLEDGEMENT

This work is supported by the National High Technology Development 863 Program of China (2006AA01Z316), the National Natural Science Foundation of China (60572137), and Science Foundation Ireland (03/IN.3/I361).

REFERENCES

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

[2] Coldefy, F., Bouthemy, P. (2004) Unsupervised Soccer Video Abstraction Based on Pitch, Dominant Color and Camera Motion Analysis. Proceedings of the 12th ACM International Conference on Multimedia, New York, NY, USA, 10-16 October, pp. 268-271.

[3] Wang, J., Chng, E, Xu, C.S., Lu, H.Q., Tian, Q. (2007) Generation of Personalized Music Sports Video Using Multimodal Cues. IEEE Transactions on Multimedia, 9(3), 576-588.

[4] Ekin, A., Tekalp, A.M., Mehrotra, R. (2003) Automatic Soccer Video Analysis and Summarization. IEEE Trans. on Image Processing, 12(7), 796-807.

[5] Intille, S. and Bobick, A. (2001) Recognizing planned, multi-person action. Comput. Vis. Image Understand., 81(3), 414-445.

[6] Tovinkere, V. and Qian, R.J. (2001) Detection semantic events in soccer games: Toward a complete solution. Proceedings of IEEE Int. Conf. Mult. Expo(ICME), Tokyo, Japan, 22-25 August, pp. 833-836.

[7] Tong, X.F., Liu, Q.S., Duan, L.Y., Lu, H.Q., Xu, C.S., Tian, Q. (2005) A Unified Framework for Semantic Shot Representation of Sports Video. Proceedings of ACM MIR'05, Singapore, 10-11 November, pp. 127-134.

[8] Bertini, M., Del Bimbo, A., Nunziati, W. (2006) Automatic Detection of Player's Identity in Soccer Videos using Faces and Text Cues. Proceedings of ACM MM'06, Santa Barbara, California, USA, 23-27 October, pp. 663-666.

[9] Ye, Q., Huang, Q., Jang, S. (2005) Jersey Number Detection in Sports Video for Athlete Identification. Proceedings of Visual Communications and Image (VCIP), Beijing, China, 12-15 July, pp. 1599-1606.

[10] Zhu, G.Y., Xu, C.S., Huang, Z.M., Gao, W., Xing, L.Y. (2006) Player Action Recognition in Broadcast Tennis Video with Applications to Semantic Analysis of Sports Game. Proceedings of ACM MM'06, Santa Barbara, California, USA, 23-27 October, pp. 431-440.

[11] Song, Y., Goncalves, L., Perona, P. (2003) Unsupervised learning of human motion. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(7), 814-827.

[12] Naidoo, W.C., Tapamo, J.R. (2006) Soccer Video Analysis by Ball, Player and Referee Tacking. Proceedings of SAICSIT 2006, Somerset West, South Africa, 9-11 October, pp. 51-60.

[13] Tong, X.F., Lu, H.Q., Liu, Q.S. (2004) An effective and fast soccer ball detection and tracking method. Proceedings of 17th International Conference on Pattern Recognition, Cambridge, UK, 23-26 August, pp. 795-798.

[14] Tang, X., Gao, X., Liu, J.Z., Zhang, H.J. (2002) A spatial-temporal approach for video caption detection and recognition. IEEE Transactions on Neural Networks, 13(4), 961-971.

[15] Pan, H., Li, B., Sezan, M.I. (2001) Detection of Slow-motion Replay Segments in Sports Video for Highlights Generation. Proceedings of IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), Salt Lake City, UT, 7-11 May, pp. 6149-1652.

[16] Ye, Q.X., Huang, Q.M., Gao, W., Jiang, S.Q. (2005) Exciting Event Detection in Broadcast Soccer Video with Mid-level Description and Incremental Learning. Proceedings of the 13th annual ACM international conference on Multimedia, 6-11 November, pp. 455-458.

[17] Leonardi, R., Migliorati, P., Prandini, M. (2004) Semantic indexing of soccer audio visual sequences: a multimodal approach based on controlled Markov chains. IEEE Trans on CSVT, 14(5), 634-643.

[18] Xie, L., Chang, S-F., Divakaran, A., Sun, H. (2002) Structure analysis of soccer video with hidden Markov models. Proceedings of IEEE Int. Conf. on Acoustics, Speech and Signal Proc.(ICASSP), Orlando, Florida, 13-17 May, pp. 4096-4099.

[19] Nepal, S., Srinivasan, U., Reynolds, G. (2001) Automatic detection of goal segments in basketball videos. Proceedings of ACM Multimedia 2001, Ottawa, Ontario, Canada, 30 Sep. - 5 Oct., pp. 261-269.

[20] Zhang, D., Ellis, D. (2001) Detecting sound events in basketball video archive. Technical Report, Dept. of Electronic Engineering, Columbia University.

[21] Ando, R., Shinoda, K., Furui, S., Mochizuki, T. (2007) A robust scene recognition system for baseball broadcast using data-driven approach. Proceedings of the 6th ACM International conference on Image and video retrieval, Amsterdam, The Netherlands, 9-11 July, pp. 186-193.

[22] Chang, P., Han, M., Gong, Y. (2002) Extract highlights from baseball game video with hidden Markov models. Proceedings of IEEE International Conference on Image Processing, Pittsburgh, PA, USA, 22-25 September, pp. 609-612.

[23] Petkovic, M., Mihajlovic, V., Jonker, M., Djordjevic-Kajan, S. (2002) Multi-modal extraction of highlights from TV formula 1 programs. Proceedings of IEEE International Conference on Multimedia and Expo, Lausanne, Switzerland, 26-29 August, pp. 817-820 .

[24] Kijak, E., Oisel, L., Gros, P. (2003) Temporal structure analysis of broadcast tennis video using hidden Markov models. Symp. Electronic Imaging: Science and Technology: Storage and Retrieval for Media Databases, 5021, 277-288.

[25] Li, B. and Sezan, M.I. (2002) Event detection and summarization in American Football broadcast video. Symp. Electronic Imaging: Science and Technology: Storage and Retrieval for Media Databases, 4676, 202-213.

[26] Peterson, J.L. (1981) Petri Net: Theory and the Modeling of Systems. Prentice-Hall, Englewood Cliffs, N.J., U.S.A.

[27] Al-Khatib, W., Ghafoor, A. (1999) An Approach for Video Meta-Data Modeling and Query Processing. Proceedings of the seventh ACM international conference on Multimedia, Orlando, Florida, USA, 30 Oct. - 5 Nov., pp. 215-224.

[28] Little, T.D.C., Ahanger, G., Folz, R.J. (1993) A Digital On-Demand Video Service Supporting Content-Based Queries. Proceedings of the First ACM international conference on Multimedia, Anaheim, California, USA, 1-6 August, pp. 427-436.

[29] Smeaton, A.F. and Gregan, A. (1994) Distributed Multimedia QoS Parameters from Presentation Modelling by Coloured Petri Nets. Lecture Notes in Computer Science; Multimedia, Hypermedia, and Virtual Reality: Models, Systems, and Applications; 1st International Conference, MHVR'94, Moscow, Russia, 14-16 September, pp. 47-60.

[30] Butz, T., Thiran, J. (2001) Shot Boundary Detection with Mutual Information. Proceedings of the IEEE International Conference on Image Processing, Thessaloniki, Greece, 10-12 Oct. 2001, pp. 422-425.

[31] Over, P., Awad, G., Kraaij, W., Smeaton, A.F. (2007) TRECVID 2007 Overview. Proceedings of TRECVid 2007 - Text REtrieval Conference TRECVid Workshop, Gaithersburg, MD, USA, 5-6 Nov. 2007.

[32] Xu, P., Xie, L., Chang, S-F., Divakaran, A., Vetro, A., Sun, H. (2001) Algorithms and System for Segmentation and Structure Analysis in Soccer Video. Proceedings of the IEEE International Conference on Multimedia and Expo, Tokyo, Japan, August 22-25 2001, pp. 721-724.

[33] Ye, Q., Huang, Q., Gao, W., Jiang, S. (2005) Exciting Event Detection in Broadcast Soccer Video with Mid-level Description and Incremental Learning. Proceedings of ACM Multimedia, Singapore, 6-11 Nov. 2005, pp. 455-458.

[34] Tjondronegoro, D., Phoebe Chen, Y-P., Pham, B. (2004) The Power of Play-Break for Automatic Detection and Browsing of Self-consumable Sport Video Highlights. Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval, New York, NY, USA, Oct 15-16 2001, pp. 267-274.

[35] Millerson, G. (1990) The Technique of Television Production (12th ed.). Focal Press, New York, USA.

[36] Harris, C.G., Stephens, M.J. (1988) A Combined Corner and Edge Detector. Proceedings of the Fourth Alvey Vision Conference, Manchester, 31 Aug. - 2 Sep., pp. 147-151.

[37] Bai, L., Hu, Y.L., Lao, S.Y., Chen, J.Y. (2005) Feature Analysis and Extraction for Audio Automatic Classification. Proceedings of IEEE System, Man and Cybernetics Conference, Hawaii, USA, 10-12 October, pp. 767-772.

[38] Bai, L., Lao, S.Y., Chen, J.Y., Wu, L.D. (2005) Audio Classification and Segmentation Based on Support Vector Machines. Journal on Computer Science (in Chinese), 4, 73-80.

[39] Bai, L., Lao, S.Y., Liao, H.X., Chen, J.Y. (2006) Audio Classification and Segmentation for Sports Video Structure Extraction Using Support Vector Machine. Proceedings of Fifth International Conference on Machine Learning and Cybernetics, DaLian, China, 13-16 August pp. 3303-3307.

[40] Jeannin, S., Divakaran, A. (2001) MPEG-7 Visual Motion Descriptors. In proceedings of IEEE Transactions on Circuits and Systems for Video Technology, 11(6), pp. 720-724.

[41] Alan F. Smeaton and Paul Over and Wessel Kraaij. (2008) High-level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. Ajay Divakaran, (eds), Multimedia Content Analysis:Theory and Applications. Springer.

[42] Gabbar, H.A. (2006) Modern Formal Methods and Applications. Springer-Verlag New York, Inc., Secaucus, NJ, USA.