Content-based organisation, analysis and retrieval of soccer video

Junqing Yu, Yunfeng He, Kai Sun and Xingming Ouyang*

School of Computer Science and Technology,
Huazhong University of Science and Technology,
Wuhan, China
E-mail: yjqing@hust.edu.cn
E-mail: yfhe@hust.edu.cn
E-mail: sunkai_xialan@163.com
E-mail: ouyangxm@hust.edu.cn
*Corresponding author

Abstract: A hierarchical organising model based-on MPEG-7 is introduced to effectively represent high-level and low-level information in soccer video. To effectively retrieve highlight clip of soccer video, adaptive abstraction is designed based on the excitement time curve, and an XML-based query scheme for semantic search and retrieval is proposed based on the hierarchical model. For retrieval, XQuery, a XML query language, is employed in the proposed querying framework. The proposed framework employs visual, audio, and cinematic features, and can be scalable to different preferences and requirements. Its efficiency, effectiveness, and the robustness of the proposed framework have been demonstrated over our extensive experiments.

Keywords: hierarchical organisation model; soccer video; event detection; semantic retrieval.

Reference to this paper should be made as follows: Yu, J., He, Y., Sun, K. and Ouyang, X. (2010) 'Content-based organisation, analysis and retrieval of soccer video', *Int. J. Computer Applications in Technology*, Vol. 38, Nos. 1/2/3, pp.64–73.

Biographical notes: Junqing Yu received his PhD in Computer Science from Wuhan University in 2002. He is currently an Associate Professor at the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China. His current research interests include digital video processing and retrieval, multimedia, and parallel algorithm.

Yunfeng He and Kai Sun are PhD Candidates of School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China. Their current research focuses on Video Information Retrieval.

Xingming Ouyang is a Professor of School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China. Her current research interest is Information Retrieval.

1 Introduction

With the rapid development of multimedia and network technologies, people can much more easily generate video content using personal camera and to access video content via network. To help users search and retrieve the interested content conveniently and effectively in the amount of video content, it is important to develop better ways to represent and organise the mass video information. Video data management and information retrieval are very important areas of research in computer technology. In recent years, many researchers from different countries have dedicated to the study of Content-Based Video Retrieval (CBVR) (Naphade et al., 2001; Deng and Manjunath, 1997; Zhang et al., 1997). However, the existing technologies of CBVR still suffer from the semantic gap between low-level features and high-level semantic concepts because computer can not directly 'calculate' the semantic meaning from low-level features, such as colour, shape, texture, motion and audio information.

For bridging the semantic gap, the intrinsic relationships between low-level features and semantic concepts should be noted, but different video genres have different useful features. For example, audio features including crowd cheers and commentator's voice can be used to explore highlights in sports video, and motion activity can be used to detect anchorperson frames in news video. Hence, a universal video description schema is more complex to catch the semantic concepts for all kinds of video, but specific-domain approach is helpful to improve the accuracy and effectiveness of the CBVR.

As an important specific-domain video genre, sports video has been widely studied for its tremendous commercial potentials. In this paper, we use specific-domain approach to represent and retrieve sports video, especially the soccer video. In the literature, researches of sports video mainly include shots boundary detection and classification, highlights extraction and events detection, as well as structure analysis and summarisation. Shot boundary detection is the first step of video analysis and can fulfilled using a set of visual features and motion features (Ekin and Tekalp, 2003). And shots classification usually uses dominant colour feature and temporal information in soccer video. Ekin classified sports video shots into long shot, in-field medium shot and close-up or out-of-field shot by dominant coloured pixel ratio (Ekin et al., 2003). Tong used three essential properties, called camera shot size, subject in a scene and video production technology to characterise a video shot, and employed the ratio of playing field and the grey-level co-occurrence matrix to detect camera shot size (Tong et al., 2005).

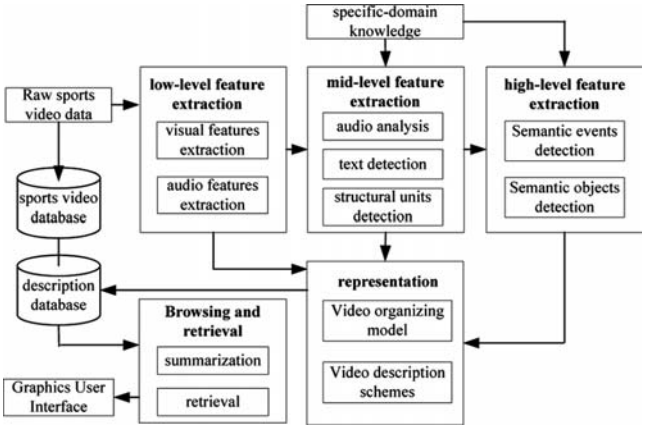
Highlights extraction and event detection is approached either by developing feature-based event models (Leonardi et al., 2003; Xu et al., 2003), by searching for keywords in speech (e.g., commentator) and closed captions (Nitta et al., 2000), by using MPEG-7 metadata (Jaimes et al., 2002) or by involving several of the above-mentioned clues into inter-modal collaboration (Snoek and Worring, 2003; Babaguchi and Nitta, 2003). The main disadvantage of these approaches can be revealed in the need for numerous and reliable event models. Clearly, it makes the event-based highlight extraction technically and semantically a complex task. The problem of different event realisations and coverage can be partly resolved via keyword spotting, the composition of the highlighting video abstract. However, this may only fit for the events like 'goal' and 'penalty' in soccer, other interesting events, such as a quick attack toward the goal finishing by a nice move of a goalkeeper in soccer, probably can not be detected.

Summarisation or abstract is usually constructed using highlight events in sports video. Ekin introduced an automatic and efficient framework using cinematic and object-based features (Ekin et al., 2003). It can output three types of summaries: all slow-motion segments in a game, all goals in a game, and slow-motion segments classified according to object-based features. Although slow-motion depicts a short time of highlight events in a game, but not the whole events. Some interesting clips like coach view after goals can not be included and some maybe repeated in summaries. Fayzullin et al. (2004) used the priority curve to create a summary of a desired length of any video. However, priorities depend on the video type and their definitions rely on different experts. In recent years, Hanjalic (2005) used motion activity, changes in shot lengths and audio energy to create a highlights time curve based on excitement modelling, and it is useful to generate hierarchical abstract. However, the highlights curve can only locate the positions of the highlight events, but can not qualify their semantics.

The high-level structure of sports video sometimes is scene, whose definition is not same in different literature. Ren defined four video segments, named play, focus, replay and break, to construct video structures called 'attack', and proposed several semantic event models based on 'attack' (Ren and Jose, 2005). Wang used a semantic concept, namely 'offense', to calculate possession and to detect goal events in soccer video (Wang et al., 2004). Kolekar used a five layered, event driven hierarchical framework for generic sports video classification (Kolekar and Sengupta, 2006). All these structure is based on shot and can be used for event detection. But they are not suitable for describing whole sports video. Therefore, a hierarchical organising model is needed to represent sports video.

In this paper, we propose an efficient framework for indexing, browsing, abstracting and retrieval of sports video. As illustrated in Figure 1, the framework consists of five stages: low-level feature extraction, mid-level feature extraction, high-level feature extraction, representation, browsing and retrieval as well. Mid-level and high-level feature extraction should use specific-domain knowledge.

Figure 1 System framework for CBVR



The rest of the paper is organised as follows. In the next section, we discuss the structure of sports videos and propose a suitable organising model and description schemes for sports video based on MPEG-7. Section 3 describes algorithms for low-level feature extraction. And highlights detection and event detection based on

excitement model are introduced in Section 4. Sports video retrieval is discussed in Section 5, with conclusion of this paper in Section 6.

2 Hierarchical organising model and description schemes for sports video

For browsing and retrieval of sports video, a sports video should be first segmented to some accessible and manageable structural units, such as shots and objects, and then use enough information, such as visual, auditory and semantic features, to describe and identify these units. In this section, an efficient hierarchical tree structure is proposed to organise the sports video and a description scheme based on MPEG-7 is introduced to represent it.

2.1 Hierarchical organising model of sports video

With the help of specific-domain knowledge, we propose a hierarchical sports video organising model as shown in Figure 2. Firstly, a sports video clip can be partitioned into several sections, which describe periods of a match, such as halves in soccer and games in volleyball. Secondly, a section is divided into several successive scenes, which depict paragraphs of a match, such as an attack in soccer games. And then, a scene is made up of several successive shots, which represent whole camera actions. Lastly, a shot is composed of some field segments, which are neighbouring match segments occurring in the same field zone and are constructed by continuous sequences of frames.

2.2 Hierarchical organising model for soccer video

Sections, scenes, shots and field segments are all the results of temporal segmentation and have different semantics in different genre of sport video. As shown in Figure 3, a hierarchical model of soccer video is used as an example to describe the semantics of these time units.

2.2.1 Section

There are two kinds of sections, namely play section and break section in soccer video. The soccer match has two halves with the equal time about 45 min besides two additional halves about 15 min in some matches which must have winner. These 2 or 4 halves are called play section. And generally, there are some time segments before, between and after halves, such as the entrance of players before a match, advertisement between halves and the celebration after a match. These time segments called break section. Generally, the time interval of a play section and the break sections are used to detect the play sections.

2.2.2 Scene

A half of soccer always includes play events, which refer to the time segments when the match is in action, and

Figure 2 A hierarchical organising model of sports video

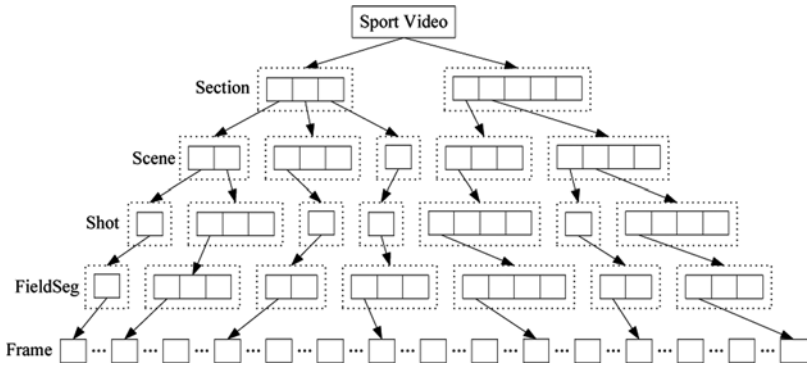
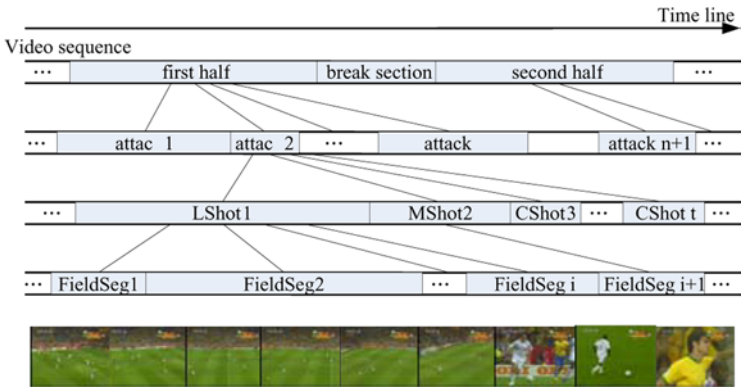


Figure 3 A hierarchical structure of soccer video (see online version for colours)



break events, which refer to the pause time in the match and represent the appearance of semantic events, such as fouls and shoots. Generally speaking, a break event describes the result of a play event before it and the start of a new play event. A scene is defined a play-break segment called 'break', which begins with a play event and ends with several break events. Scene detection usually uses the shot type, shot length and some domain knowledge (Ekin and Tekalp, 2003). For example, in soccer video, long time global view shots and middle view shots indicate a play event, while close-up shots usually induce a break event.

2.2.3 Shot

In soccer video, shots usually can be classified to long shots, medium shots, close-up shots and other shots. Long shot display a global view of the playfield, medium shot is a zoomed-in view of a specific part of the field and usually includes the whole body of players, and close-up shot shows the above-waist view of a player, while some out-of-field shots in soccer video like audience view belong to other shots (Ekin et al., 2003).

2.2.4 Field segment

In some universal organising model, shots are the basic representing and accessing units of the video content. But in sports video, a long shot always lasts a long time, longer than 100 frames, and sometimes includes play events, which are boring or have no results. Therefore, shots should be divided into several field segments to describe play events occurring in a specific part of field. A soccer field is divided into 12 zones, six for each side (Assfalg et al., 2003). Identification of the field zones is helpful to detect highlight events, such as shoot events and corner kick events. The algorithm of field segments detection is based on the detection of field using dominate colour feature. Generally, a long shot can be divided into several field segments, while a medium shot or a close-up shot does not need to be partitioned whatever the court zone is detected or not.

2.3 Description schemes based-on MPEG-7

MPEG-7 standard, formally known as Multimedia Content Description Interface, is an ISO/IEC standard developed by MPEG. Multimedia Description Schemes (MDS), one of the main parts of MPEG-7, provides a rich set of standardised description tools called Descriptors (Ds) and Descriptor Schemes (DSs) to describe many aspects of multimedia. The multimedia content description tools of MDS are divided into structure descriptions tools and concept description tools, which describe the structure and semantics of multimedia content respectively (Salembier and Smith, 2001). MPEG-7 also provided a Description Definition Language (DDL), which is based on the XML Schema Language to define new Ds and DSs. In this section, we will propose our description schemes based on the generic description schemes provided by MPEG-7, DDL and specific-domain acknowledge enables more efficient retrieval.

In our description schemes, VideoDS, SectionDS, SceneDS, ShotDS and FieldSegD, are used to describe the content units of soccer video, including sections, scenes, shots and field segments. Every content unit is described by four aspects: attributes, structure partition, low-level features and semantic features.

2.3.1 VideoDS

VideoDS identifies video clip using attributes, structure partition and semantic features. It can be used to describe the whole sports video. 'VideoID' is an ID attribute used to identify a video. VideoDS has two other attributes: genre and type. Genre defines the specific video, such as sports, news and movies etc. And the type attribute define the type of a specific domain video, such as soccer, basketball and baseball etc for the sports video. VideoDS also includes metadata, some of which can not be acquired automatically and should be inputted manually. In soccer video, metadata usually include media location, media time, match name, match place, match score, as well as the names, colour and player lists of both host and visiting team.

2.3.2 SectionDS

SectionDS identifies sections in the sports video. There are two attributes, 'SectionID' and 'type', and two metadata, 'MediaTime' and 'SectionScore'. 'SectionID' attribute is used to identify a section. 'Type' attribute uses a binary value to represent the section type, '0' for play section and '1' for break section respectively. 'MediaTime' element represents the start time-point and duration time of the section. 'SectionScore' is used to record the score of a section. And a section is partitioned into several scenes, whose relationships can be described using RelationGraphDS tool of MPEG-7.

2.3.3 SceneDS

SceneDS identifies a scene, which is also a semantic unit of video. 'SceneID' attribute is used to identify a scene. A scene usually includes one or more semantic events, which are described by SEventDS based on the EventDS of MPEG-7. Because the event time-point in video is different from the one in game, so SEventDS includes two time attributes identifying event time-point in video and in game respectively. A scene also includes several shots. The relation between events and the one between shots in a scene are described using RelationGraphDS.

2.3.4 ShotDS and FieldSegDS

ShotDS and FieldSegDS are used to identify accessible units, which are described by low-level features and semantic features. The low-level features include visual, audio and motion features, while semantic features mainly consist of objects and events dependent on the type of the shot or field segment. For example, close-up

shots usually depict an object, which may be a player, a referee or a coach. While a medium shot or field segment depict an event. A 'type' attribute uses '0' to represent long shot, '1' for medium shot and '2' for close-up shot. A 'FieldType' element, which has four attributes: shape, size, corner and midfield, are used to describe the type of shot and field segment respectively. Especially, in order to easily construct video abstract in different granularity, a 'HighlightRank' attribute, whose value is a float from 0.0 to 1.0, is used in ShotDS and FieldSegDS to represent the level of excitement.

Using the above description schemes, the content of sports video can be identified and retrieved effectively. An example has given to show how these schemes are instantiated in Figure 4.

Figure 4 A description example of soccer video

```
<Video genre="sport" type="soccer" VideoID="soccer_1">
 <MetaDescriptor>
  <MediaLocator>
   <MediaUri>http://SportsVideoLibrary.org/soccer1.mpg</MediaUri>
  </MediaLocator>
  <MatchName>FIFA World Cup</MatchName>
  <MatchPlace>Frankfurt, Germany</MatchPlace>
  <MatchTime>2006-7-2T3:00</MatchTime>
  <HostTeam><Name>Brazil</Name>…</HostTeam>
  <VisitingTeam><Name>France</Name>…</VisitingTeam>
  <FinalScore>H0-V1</FinalScore>
 </MetaDescriptor>
 <section SectionID="break_1" type="0">
   …
 </section>
 <section SectionID="half_1" type="1">
  <MetaDescriptor>
   <MediaTime>
    <MediaRelTimePoint>PT10M15S27F100</MediaRelTimePoint>
    <MediaDuration>PT44M41S71F100</MediaDuration>
   </MediaTime>
   <SectionScore>H0-V0</SectionScore>
  </MetaDescriptor>
  <scene SceneID="Scene_1">
   <MetaDescriptor>…</MetaDescriptor>
   <event id="foul_1">
    <Label><Name>Foul</Name></Label>
    <SemanticTime>PT10M46S14F100</SemanticTime>
    <GameTime>PT16S</GameTime>
   </event>
   <shot ShotID="shot_1" type="0">
    <MetaDescriptor>…</MetaDescriptor>
    <HighlightRank>0.2</HighlightRank>
    <fieldseg FieldSegID="seg_1">
     <MetaDescriptor>…</MetaDescriptor>
     <HighlightRank>0.1</HighlightRank>
     <FieldType>
      <Shape>0</Shape>  <Size>0.62</Size>
      <Corner>0</Corner> <MidField>1</MidField>
     </FieldType>
    </fieldseg>
    ...
</video>
```

3 Feature extraction

Feature extraction plays a vital role in content-based video analysis. The features are the information extracted from source video. They are represented in a suitable way, stored in an index, and used for analysis and retrieval. Initially, researchers focus on visual features such as colour, texture, shape and motion. More recently, on the basis of media archive feedback, researchers have emphasised on the combination audio and visual features.

Features listed in MPEG-7 include audio and visual features. The visual features are composed of colour, texture, shape and motion. The audio ones contain frame and segment features, such as silence, frequency centroid, power, timbre and signature etc. More details of the features in MPEG-7 can be referred to ISO/IEC/JTC1/SC29/WG11 (MPEG) (2001a, 2001b). Here, we just discuss the features different from those in MPEG-7.

3.1 Visual features

Motion Activity (*MA*): The motion activity, defined as a total motion in the scene including both the object and camera motion. The motion activity, MA(k) at video frame k, can be computed as the average magnitudes of all motion vectors in the frame. Using a block-based motion estimator, the motion vector can be computed. Then we can normalise the motion vector with the maximum motion value $|\vec{\nu}_{\max}|$

$$\mathrm{MA}(k) = \frac{\left[\sum_{j=1}^{B} |\vec{\nu}_j(k)|\right]}{B|\vec{\nu}_{\max}|} \tag{1}$$

where B is the number of blocks within a frame and $\vec{\nu}_j(k)$ is the motion vector of the block j in frame k.

Shot Density (*SD*): In order to measure the density of shots, we defined the shot density function SD(k):

$$\mathrm{SD}(k) = \mathrm{e}^{\left(\frac{1-(n(k)-p(k))}{\delta}\right)}. \tag{2}$$

Here, $p(k)$ and $n(k)$ are the positions (frame indexes) of the two adjacent shots of the frame k. The parameter δ is a constant determining the way, in which the SD(k) values are distributed on the scale between 0 and 1.

3.2 Audio features

Sound energy in higher frequencies: The sound energy is computed for each video frame. The audio samples that cover the same time period as one video frame is $s = F/f$, where f is the video frame rate and F is the audio sampling frequency (typically 44.1 KHz for CD quality). The power spectrum is computed for each consecutive segment of the audio signal containing s samples. The short time energy, $STE(k)$, is then computed by summing up all spectral values from a predefined cut-off frequency C.

4 Scalable soccer video abstracting and event recognition based on excitement modelling

As a spin-off from the previous work of Hanjalic and Xu (2005), we propose a method for scalable

soccer video abstracting that is based on modelling the expected variations in the excitement level of the audience. Since it is realistic to assume that a highlighting event (e.g., goal, red/yellow card, foul) induces a steady increase in an audience's excitement, we search for highlights in those video segments that are expected to excite the audience most. By grouping these highlights and tuning the excitement level parameter, the scalable soccer video abstract can be generated automatically.

4.1 Excitement modelling

As 'excitement' is a psychological category, we first introduce two basic criteria. The first criterion – *Comparability* – ensures that excitement levels obtained in different sequences for similar types of events are comparable. This criterion obviously imposes normalisation and scaling requirements when computing the excitement time curve. The second criterion – *Smoothness* – accounts for the degree of memory retention of preceding frames and shots. This criterion was also adopted by Adams et al. (2000), who constructed a continuous function illustrating the changes in tempo/pace along a movie. The smoothness criterion takes the perception of the content into account since the affective state of audience does not change abruptly from one video frame to another.

4.1.1 Feature selection

A number of psycho-physiological studies have been performed concerning the effect of non-content (structural) attributes of film and television messages on the affective state of the audience. One of the most extensively investigated attributes is motion. Research results show that motion in a television picture seems to have a significant impact on individual affective responses. This has been also realised by film theorists who contend that motion is highly expressive and is able to evoke strong emotional response in audience. Specially, Detenber et al. (1997) investigated the influence of camera and object motion on emotional responses of the tested persons and concluded that an increase in motion intensity on the screen causes an increase in audience's arousal.

Further, as has been shown by Picard (1997), various characteristics of the audio and/or speech stream of a program seem to provide valuable clues about the affective content of that program. Specially, the loudness (signal energy) is known to be directly related to the evoked level of audience's excitement.

Beside low-level audiovisual features, some editing effects are also very useful to infer the values of excitement. The density of cuts is one good example of such effects. As also realised by Adams et al. (2000), cuts are a popular tool for the director to either create the desired pace of action (e.g., in a movie) or to respond to interesting events in live broadcasts (e.g., goals in a soccer game). In order to maximise the desired effect, the director deliberately uses cuts instead of some other editing effects, such as dissolves, fades or wipes, which spread in time and are used in more stationary (i.e., less exciting) program. By varying the cut density, a movie director establishes the relationship between the action and the audience's attention. In this sense, it is justifiable to link the varying cut density to the level of excitement induced in audience during a movie. The relation between the cut density and excitement becomes even clearer in live broadcasts of sport events. It can be explained by an example of a soccer match. If there are no highlight events, it is broadcasted by one camera that covers the entire field and follows the game in one continuous shot. However, whenever there is a goal or other highlight events, the director immediately increases the density of cuts trying to show the details in the field or among the spectators at that moment from different directions. The cut density can also increase when there is an important break (e.g., foul play, free kick, etc.). Obviously, any increase in cut density during such broadcasts is the reaction of a director to an increase in the excitement evoked in the sport arena.

Motivated by the above observation, we model the excitement as a function of the following components:

1 overall motion activity measured at frame transitions (MA)

2 the energy in the audio track of a video program (STE)

3 the density of cuts (SD).

4.1.2 Excitement modelling and excitement time curve

Our excitement model is based on the arousal model proposed in Kobla et al. (2000). The curve $A(k)$ was defined as a function of N basic components $G_i(k)$

$$A(k) = F(G_i(k),\ i = 1, \ldots, N). \tag{3}$$

To simplify the terminology, we will refer to $A(k)$ in this paper as *excitement time curve*. The function $G_i(k)$ models the variations in the excitement level over the video frame k induced by the stimulus represented by the feature i. It is assumed that all of the N features were selected to reliably represent the stimuli which influence the affective state of the audience. Each function $G_i(k)$ can be seen as an elementary excitement time curve or the primitive of the overall excitement time curve $A(k)$, while $A(k)$ is obtained by integrating the contributions of all component $G_i(k)$ using a suitable function F.

Based on the discussion in Section 4.1.1, here $N = 3$. The $G_1(k)$, $G_2(k)$ and $G_3(k)$ can be computed as follows.

$$\begin{aligned} G_1(k) &= \frac{\max_k(MA(k))}{\max_k(\widetilde{MA}(k))}\widetilde{MA}(k), \\ G_2(k) &= \frac{\max_k(SD(k))}{\max_k(\widetilde{SD}(k))}\widetilde{SD}(k), \\ G_3(k) &= STE'(k)(1 - \overline{STE}). \end{aligned} \tag{4}$$

Here $\widetilde{MA}(k)$ is the result of the convolution of the curve $MA(k)$ with a smoothing window, that is $\widetilde{MA}(k) = MA(k) \times K(l_1, \beta_1)(K(l_1, \beta_1)$ is the Kaiser window of the length l_1 and the shape parameter β_1), $\widetilde{SD}(k) = SD(k) \times K(l_1, \beta_1)$. The $STE'(k)$ and $\overline{STE}$ in the $G_3(k)$ are defined as

$$STE'(k) = \frac{\widetilde{STE}(k)}{\max(\widetilde{STE}(k))} \tag{5}$$

and

$$\overline{STE} = \frac{1}{W}\sum_k STE'(k), \tag{6}$$

$$\widetilde{STE}(k) = STE(k) \times K(l_1, \beta_1) \tag{7}$$

(W is the total frame number of the tested video).

Figure 5(a) shows the three arousal components computed for an excerpt from a typical soccer video in *Word Cup 2006*. It can be seen that local maxima exist in the three curves when exciting events (goals, chances) occurs. One can also notice that these local maxima are not necessarily aligned. For instance, in the case of a goal, the following scenario is possible: the audience first cheer the action (sound energy peak), then there are cameras zooming to running players (motion activity peak) and, finally, there are cameras zooming to the teams' coach or audience (cut density peak). This fact motivates the definition of the function F as a weighted average of the three components, which is then convolved with a sufficiently long Kaiser window in order to merge neighbouring local maxima of the components. The result is finally re-scaled to the 0–1 range. The process is shown in formula (8)

$$A(k) = \frac{\max_k(a(k))}{\max_k(\tilde{a}(k))}\tilde{a}(k) \tag{8}$$

with

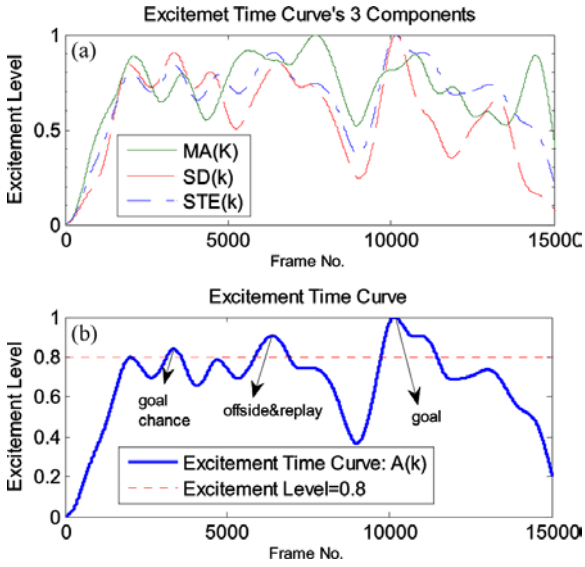
$$\begin{aligned} a(k) &= \frac{1}{k}\sum_i G_i(k), \\ \tilde{a}(k) &= a(k) \times K(l_2, \beta_2). \end{aligned} \tag{9}$$

The excitement time curve computed by formula (8) is demonstrated in Figure 5(b). The length and shape parameter of the Kaiser window used to smooth the excitement time curve were set to 2000 and 5, respectively.

4.2 Generation of the scalable soccer video abstract

Given the excitement time curve $A(k)$ and the desired maximum abstract length L in frames, highlights can be extracted by analysing the values of the curve and those video segments that are likely to excite the audience most can be extracted. To do this, we can set a constant value and draw a horizontal line cutting off the peaks of the curve in Figure 5(b). The video segments between the pairs of intersection points are chose as the content of video abstract. The Scalable video abstract can be obtained by adjusting the excitement level. As the extraction process is driven by the local excitement level only, any event in a soccer video may be included into the abstract, provided that the curve passes through a sufficiently high value range during that event. In this way, highlights are extracted in a generic fashion without the need for event modelling or artificially limiting the scope of the abstract content.

Figure 5 (a) Three component time curves and (b) corresponding excitement time curve obtained for an excerpt from a soccer match in World Cup 2006 (see online version for colours)



The red horizontal line in Figure 5(b) provides an abstract of about 2 min (the total excerpted soccer video is 10 min), showing a goal chance, an offside and a goal contained in the analysed excerpt. The effect of our scalable soccer video abstracting method becomes clear when we move the cutoff line vertically.

4.3 Soccer event recognition

This section presents algorithms for the detection of soccer video events within the generated abstract upon the excitement time curve, such as replays and goals. Specifically, we develop a simple but very effective replay detection algorithm that uses MA feature and a soccer event detection algorithm that uses only cinematic features (domain knowledge) based on the excitement time curve.

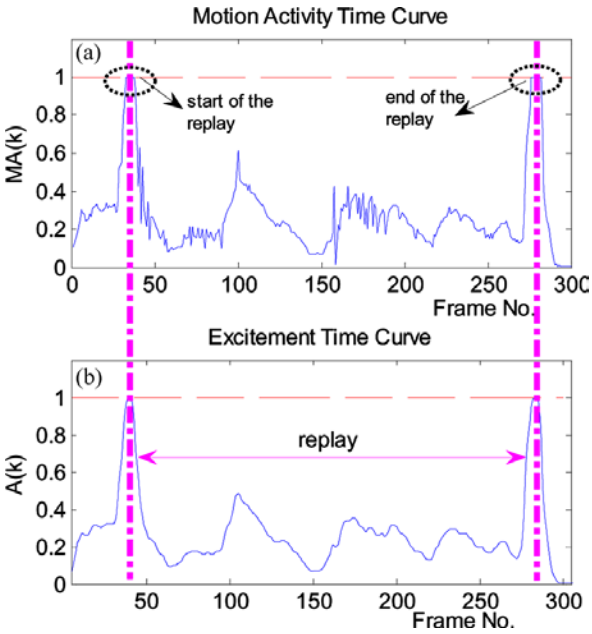
4.3.1 Replay detection

Replay in soccer video is an excellent locator of semantical segment for high level video analysis. Several slow-motion replay detectors for compressed (Kobla et al., 2000) and spatial domain (Pan et al., 2002) exist in the literature. Here, we use the broadcaster-dependent

logo transitions, which are used at the starting and ending of replays, to detect them.

From Figure 6(a) we can find that when a logo transition occurs, the value of $MA(k)$ will be increased dramatically. In order to mark these locations, we filter the $MA(k)$ time curve by a threshold Th. That is, if $MA(k) > Th, MA(k) = 1$; otherwise, $MA(k) = 0$. So the neighbour values of filtered $MA(k)$ at the location of a logo transition will be several successive one (e.g., $\ldots 000\underline{11111}000\ldots$). Thus, the replay in a shot can be recognised easily by checking the existence of sequences composed by one in the filtered $MA(k)$. If the total number of one is bigger than a predefined number (e.g., 5), there may be a start of a replay. If the same condition is satisfied at the end of the neighbour shots, we can assert that these shots must within a replay. Combined with excitement time curve, we can get more accurate detection result. The main idea of this algorithm is demonstrated in Figure 6.

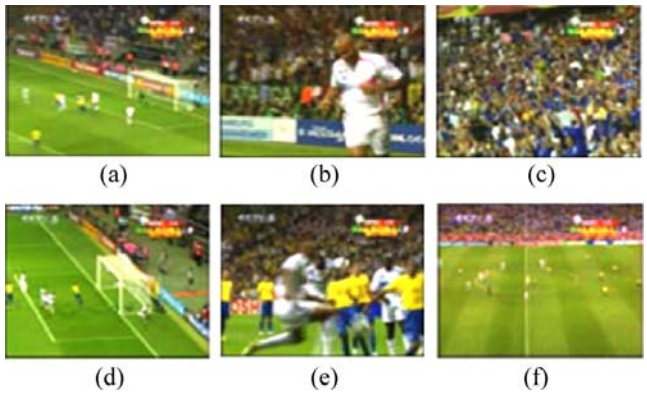
Figure 6 (a) MA (k) time curves and (b) corresponding excitement time curve obtained for an excerpt from a soccer match in World Cup 2006 (see online version for colours)



4.3.2 *Goal event detection*

A goal is scored when the whole of the ball passes over the goal line between the goal posts and under the crossbar. However, it is difficult to verify these conditions automatically and reliably by the state-of-the-art video processing algorithms (Ekin et al., 2003). Instead, we exploit our observation about the broadcasters consistent use of a certain pattern of cinematic features after the goal events. The occurrence of a goal event leads to a break in the game. During this break, the producers convey the emotions in the field to the TV audience and show one or more replay(s) for a better visual experience. The emotions are captured by one or more close-up views of the actors of the goal event, such as the scorer and the goalie, and by shots of the audience celebrating the goal. Furthermore, several slow-motion replays of the goal event from different camera positions are shown. Finally, the restart of the game is usually captured by a long shot. Figure 7 illustrates the instantiation of such a pattern.

Figure 7 An example for cinematic goal template (the temporal order is from (a) to (f)): (a) long view of the actual goal play; (b) player close-up; (c) audience; (d) the first replay; (e) the second replay and (f) long view of the start of the new play (see online version for colours)



Other interesting events may also fit this template, although not as consistently as goals. The addition of such segments into the video abstract may even be desirable since each such segment consists of interesting segments. Therefore, the recall rate for this algorithm is much more important than the precision rate, since the users will not be tolerant to missing goals, but may enjoy watching interesting non-goal events.

We define the following rules for the *cinematic goal template* that occurs between the long shot resulting in the goal event and the long shot that shows the restart of the game:

- *Duration of the break*: A break that is due to a goal lasts no less than 30 and no more than 120 s.
- *The occurrence of at least one close-up/out-of-field shot*: This shot may either be a close-up of a player or out-of-field view of the audience.
- *The existence of at least one replay shot*: The goal play is always replayed one or more times.
- *The relative position of the replay shot*: The replay shot(s) follow the close-up/out-of-field shot(s).

In order to locate the template positions, we use excitement time curve. For every detected peek of the curve (e.g., the three peeks in Figure 5(b)), the system finds the long shots that define the start and the end of the corresponding break. Finally, the rules, defined above, are verified to detect goals.

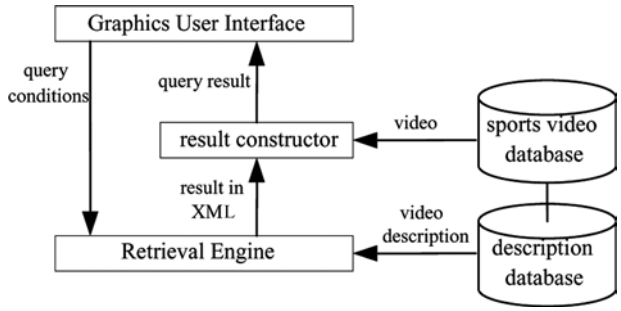
5 Semantic retrieval of sports video

5.1 Retrieval tools

Semantic retrieval plays a kernel role in CBVR, and the architecture of semantic retrieval in our system is shown in Figure 8. There are three parts in the architecture.

- *Graphics User Interface* (*GUI*). GUI accepts query conditions inputted by users and turns these conditions with other necessary information, such as the bandwidth of network, to retrieval engine. Furthermore, GUI shows the query results to users.
- *Retrieval engine.* Retrieval engine firstly gets query conditions from GUI, and then analyses the query and constructs a syntax tree for the query using a parser. Secondly, retrieval engine retrieves video description documents in description database using the syntax tree and produces a result in XML document.
- Result constructor. The result constructor analyses the result in XML, then extracts relevant content from video database and combines these contents using video processing algorithms. Lastly, the result constructor sends the content with the result in XML to GUI.

Figure 8 Architecture of the semantic retrieval



5.2 Retrieval language

Since our description schemes are defined using DDL based on XML schema, XML query languages can be used in our system. Xquery is a forthcoming standard for querying XML data proposed by W3C, and it borrows features from several query languages, including Xpath, XQL, XML_QL, SQL and OQL (Boag et al., 2006).

Xquery is able to combine information from different parts of a XML document or multiple XML documents and supports all operations on all data types represented by the data model. Additionally, it also supports hierarchies and sequence of document structures (Tjondronegoro and Phoebe Chen, 2002). For these reasons, Xquery is suitable for out system to query description data.

There are two query approaches provided by Xquery, namely path expression and FLWOR (For, Let, Where, Order by, and Return) expression.

Path expression based on Xpath is suitable for querying tree structure. Path expression consists of several steps, and includes one or more query conditions. The result of a path expression is a node list meeting the query conditions. For example, we can use a path expression as the follows to query shots with highlight rank more than 0.8 in a soccer video for summarisation.

```
//video[@genre =“sport”][@type=“soccer”]
/section/scene/shot [./HighLightRank > = 0.8]
```

FLWOR expression looks more like SQL. The ‘for’ and ‘let’ clauses expression generate an ordered the tuple stream, which is a sequence of tuples of bound variables. The optional ‘where’ clause serves to filter the tuple stream. The optional ‘order by’ clause is used to reorder the tuple stream. And the ‘return’ clause constructs the result of the FLWOR expression, using the variable bindings in the respective tuples (Tjondronegoro and Phoebe Chen, 2002). FLWOR expression is more complex than path expression, but the result of FLWOR expression is more flexible. For example as follows, a FLWOR expression does the same query as before, but only returns the ‘MediaTime’ of ‘shot’ node and uses a ‘ShotList’ node including the result.

```
<ShotList>{
  for $s in
//video[@genre=“sport”][@type=“soccer”]/
section/scene/shot
  where $s/HighlightRank>= 0.8
  return
<shot>{$s/MetaDescriptor/MediaTime}</shot>
 }</ShotList>
```

6 Conclusion and future work

In this paper, a novel content-based soccer video retrieval framework is presented. A hierarchical organising model and description schemes based on MPEG-7 are used to describe soccer video for indexing, browsing and summarisation effectively. And some important algorithms of video processing are discussed in details.

Using specific-domain approaches, video analysis and semantic extraction become easy. Therefore, our framework is useful to bridge the semantic gap. The topics for future work includes:

- combination of visual, audio and textual features to increase the accuracy of shot boundary detection and event detection
- extension of the proposed approaches to different sports, such as basketball, volleyball, and baseball
- implementation of retrieval framework and human-centered GUI.

Acknowledgement

We gratefully acknowledge the supports received from the National Science Foundation of China (60703049).

References

Adams, B., Doni, C. and Venkatesh, S. (2000) 'Novel approach to determining tempo and dramatic story sections in motion pictures', *IEEE International Conf. on Image Processing*, Vancouver, Canada, pp.13–16.

Assfalg, J., Bertini, M., Colombo, C., DelBimo, A. and Nunziati, W. (2003) 'Automatic extraction and annotation of soccer video highlights', *ACM International Conference on Image Processing*, Melbourne, USA, pp.527–530.

Babaguchi, N. and Nitta, N. (2003) 'Intermodal collaboration: a strategy for semantic content analysis for broadcasted sports video', *Proc. IEEE ICIP 2003*, Barcelona, Spain, September, pp.13–16.

Boag, S., Chamberlin, D., Fernandez, M., Florescu, D., Robie, J. and Siméon, J. (2006) 'Xquery 1.0: an XML query language', *W3C Candidate Recommendation*, June, pp.5–25.

Deng, Y. and Manjunath, B.S. (1997) 'Content based search of video using color, texture and motion', *Proc. IEEE Int. Conf. Image Processing*, Santa Barbara, CA, Vol. 2, October, pp.13–16.

Detenber, B.H., Simonr, R.F. and Bennelt, G.G. (1997) "Roll'em!: the effects of picture motion on emotional responses', *Journal of Broadcasting and Electronic Media*, Vol. 21, pp.112–126.

Ekin, A. and Tekalp, A.M. (2003) 'Generic play-break event detection for summarization and hierarchical sports video analysis', *IEEE International Conference on Multimedia and Expo ICME '03*, Baltimore, USA, Vol. 1, Nos. 6–9, July, pp.169–172.

Ekin, A., Tekalp, A.M. and Mehrotra, R. (2003) 'Automatic soccer video analysis and summarization', *IEEE Transactions on Image Processing*, Vol. 12, No. 8, July, pp.796–807.

Fayzullin, M., Subrahmanian, V.S., Albanese, M. and Picariello, A. (2004) 'The priority curve algorithm for video summarization', *Proceedings of the 2nd ACM International Workshop on Multimedia Databases*, November, Washington, USA, pp.28–35.

Hanjalic, A. (2005) 'Adaptive extraction of highlights from a sport video based on excitement modeling', *IEEE Transactions on Multimedia*, Vol. 7, No. 6, December, pp.1114–1122.

Hanjalic, A. and Xu, L-Q. (2005) 'Affective video content representation and modeling', *IEEE Trans on Multimedia*, No. 1, Vol. 7, February, pp.143–154.

ISO/IEC/JTC1/SC29/WG11 (MPEG) (2001a) *Text of ISO/IEC 15938-3 Multimedia Content Description Interface – Part 3: Visual. Final Committee Draft*, Document No. N4062, Singapore, March.

ISO/IEC JTC1/SC29/WG11 (MPEG) (2001b) *Text of ISO/IEC 15938-3 Multimedia Content Description Interface – Part 4: Audio. Final Committee Draft*, June.

Jaimes, A., Echigo, T., Teraguchi, M. and Satoh, F. (2002) 'Learning personalized video highlights from detailed MPEG-7 metadata', *Proc. IEEE ICIP 2002*, Rochester, NY, Vol. 1, pp.133–136.

Kobla, V., DeMenthon, D. and Doermann, D. (2000) 'Identifying sports videos using replay, text, and camera motion features', *Symp. Electronic Imaging: Science and Technology: Storage and Retrieval for Media Databases*, San Jose, USA, Vol. 3972, January, pp.332–343.

Kolekar, M.H. and Sengupta, S. (2006) 'A hierarchical framework for generic sports video classification', *7th Asian Conference on Computer Vision*, Hyderabad, India, January, pp.633–642.

Leonardi, R., Megliorati, P. and Prandini, M. (2003) 'Semantic indexing of sports program sequences by audio-visual analysis', *Proc. IEEE ICIP 2003*, September, Barcelona, Spain, pp.9–12.

Naphade, M.R., Wang, R. and Huang, T.S. (2001) 'Multimodal pattern matching for audio-visual query and retrieval', *Proceedings of the Storage and Retrieval for Media Datbases*, San Jose, USA, Vol. 4315, pp.188–195.

Nitta, N., Babaguchi, N. and Kitahashi, T. (2000) 'Extracting actors, actions and events from sports video – a fundamental approach to story tracking', *Proc. 15th Int. Conf. Pattern Recognition (ICPR 2000)*, Barcelona, Spain, Vol. 4, pp.718–721.

Pan, H., Li, B. and Sezan, M.I. (2002) 'Automatic detection of replay segments in broadcast sports programs by detection of logos in scene transitions', *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Orlando, Florida, pp.3385–3388.

Picard, R. (1997) *Affective Computing*, MIT Press, Cambridge, MA, pp.21–45.

Ren, R. and Jose, J.M. (2005) 'Football video segmentation based on video production strategy', *27th European Conference on IR Research ECIR2005*, Santiago, Spain, March, pp.433–446.

Salembier, P. and Smith, J.R. (2001) 'MPEG-7 multimedia description schemes', *IEEE Transactions on Circuits and Systems for Video Technology,* Vol. 11, No. 6, June, pp.748–759.

Snoek, C.G.M. and Worring, M. (2003) 'Time interval maximum entropy basedevent indexing in soccer video', *Proc. IEEE Int. Conf. Multimedia and Expo 2003 (ICME '03)*, Baltimore, USA, Vol. 3, pp.481–484.

Tong, X., Liu, Q., Duan, L., Lu, H., Xu, C. and Tian, Q. (2005) 'A unified framework for semantic shot representation of sports video', *Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval*, Singapore, November, pp.127–134.

Tjondronegoro, D. and Phoebe Chen, Y-P. (2002) 'Content-based indexing and retrieval using MPEG-7 and X-Query in video data management systems', *World Wide Web: Internet and Web Information Systems*, Vol. 5, No. 3, pp.207–227.

Wang, L., Lew, M. and Xu, G. (2004) 'Offense base temporal segmentation for event detection in soccer video', *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval*, New York, October, pp.259–266.

Xu, G., Ma, Y-F., Zhang, H-J. and Yang, S. (2003) 'A HMM based semantic analysis framework for sports game event detection', *Proc. IEEE ICIP 2003*, September, Barcelona, Spain, pp.25–28.

Zhang, H., Wang, A. and Altunbasak, Y. (1997) 'Content-based video retrieval and compression: a unified solution', *Proc. IEEE Int. Conf. Image Processing*, Santa Barbara, CA, Vol. 1, October, pp.13–16.