6

Semantic Based Sport Video Browsing

Xueming Qian
School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China

1. Introduction

In this chapter, we focus our attention on semantic based sport video highlights detection and semantic based sport video browsing. Users are more interested in sport video highlights than the normal kicks. They site before their TV sets to enjoy the exciting moments that their favorite teams shooting goals. As audiences enjoy the highlight of the video content, usually they have patient to the inserted video Advertisements. For the web based video services, the click rates of sport video highlights are far more than the whole video sequences [56]. Detecting highlights effectively is not only of interest to users/audiences (they are willing to view the highlights) but also of interest to commercial companies. They are interested in inserting their advertisement around the highlights to get more attention from consumers and expand the influences of their products/services. In many online video services websites, the sport video highlights are manually labeled which is very time-consuming and expensive. Thus automatic sport video highlight detection is very urgent [1]-[29]. It is a fundamental step for semantic based video browsing [1,7,8]. However, due to the semantic gaps between computer and human beings, highlights detection is not a trivial.

Two types of approaches have been widely utilized in sport video highlight detection, as shown in Fig.1. The first type of highlights detection approaches are carried out by mapping from low-level features directly or using model-based approaches. In the second type of approaches a mid-level semantic layer is introduced between the low-level features and high-level semantics. Highlights are detected from mid-level semantics rather than from low-level features directly. Thus it is robust against the divergences of low-level features. The second type of sport video highlight detection approaches is more effective than the first type of approaches. Based on the detected high-level semantics, semantics based video browsing can be performed.

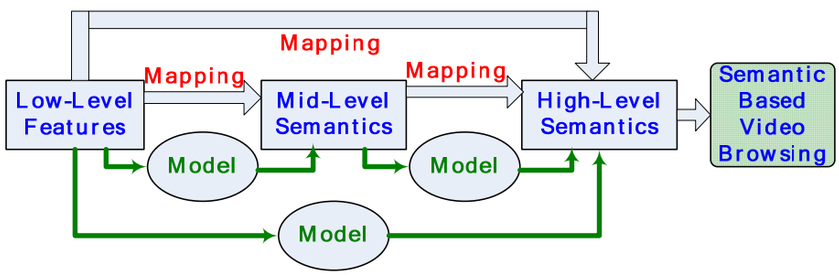


Fig. 1. Block diagram of semantic based sport video browsing.

The conventional video content browsing is based on the summarization of the whole video content. However viewers may be interested to the highlights of sport video. In this chapter, a semantic based sport video browsing approach is introduced. It is a novel book style based approach. Sport video summarization structure is similar to the table of content (ToC) of books. The semantic based video browsing approach has several advantages. Firstly, it is a hierarchical video summarization framework. Secondly, it provides seamless connections with sport video high-level semantics extraction. Thirdly, it is very convenient for users to find their interested content in a large scale database.

In this chapter, firstly learning based soccer video highlight detection approaches are expressed and then semantic based soccer video browsing approach is introduced. The main content of this chapter is organized as follows. Effective low level features representation for sports video is given in Section 2. Middle level semantic classification approaches for soccer video is provided in Section 3. High level semantic detection approaches are introduced in Section 4. Semantic based video browsing is given in Section 5 and conclusions are drawn in Section 6.

2. Low-level feature representation

In this section, several effective low-level visual features for sport video content analysis are introduced. The features are dominant color, motion, texture and overlaid text information. From the low-level visual feature detection results, mid-level semantics classification and high-level events can be determined by using either direct mapping or statistical learning based approaches.

2.1 Dominant color

Dominant color is an effective feature for soccer video analysis. The dominant color is essentially the color of the grass field. The distribution and percentage of grass field region offer significant cues for mid-level semantics categorization and highlights detection.

2.1.1 Related work on dominant color extraction

In [14], Duan et al. classified video shots into eight categories by fusing the global motion pattern, color, texture, shape, and shot length information in a supervised learning framework. Motion and dominant color information of a video shot are fused by multi-layer hidden Markov models (HMM) to determine its categorization [6]. Dominant color ratio and the dominant color projection histogram [12] are utilized for semantic soccer video shot classification. The dominant color extraction approach consists of two stages: dominant color modeling and adaptive dominant color refinement [12]. The dominant color modeling can be viewed as determining a coarse dominant color, which can be utilized for various soccer videos. While the dominant color refinement can be viewed as getting more accurate dominant color for a specific video.

In [56], dominant color detection scheme consists of three steps: 1) Initial dominant color modeling in HSI color space. The accumulative histograms of the HSI components are constructed from training frames of global views randomly selected form wide range of soccer videos and the initial dominant color are determined from the color histogram [36]. 2) Initial dominant color region determination. The initial dominant color and the cylindrical

metric [36] are then utilized to classify each pixel of current frame into dominant color or non-dominant color. 3) Adaptive dominant color refinement. For the pixels labeled as initial dominant color, their accumulative histograms of HSI are reconstructed again and the same operations as in initial dominant color modeling step are utilized to obtain the refined dominant color (H_0, S_0, I_0) of the current frame. Let's give a brief overview of this approach.

2.1.2 Coarse to fine dominant color extraction approach

The distance of a pixel located at coordinate (*i*, *j*) with color $(H(i,j), S(i,j), I(i,j))$ to the dominant color (H_0, S_0, I_0) is measured by the cylindrical metric [36] as follows

$$D(i,j) = \sqrt{D_{int}(i,j)^2 + D_{chr}(i,j)^2} \tag{1}$$

where $D_{int}(i,j)$ and $D_{chr}(i,j)$ denote the distance in intensity and chrominance components respectively.

$$D_{int}(i,j) = I(i,j) - I_0 \tag{2}$$

$$D_{chr}(i,j) = \sqrt{S(i,j)^2 + S_0^2 - 2S_0 S(i,j)\cos(\theta(i,j))} \tag{3}$$

$$\theta(i,j) = \begin{cases} |H(i,j) - H_0| & if\ |H(i,j) - H_0| \le 180^\circ \\ 360^\circ - |H(i,j) - H_0| & otherwise \end{cases} \tag{4}$$

Similar to [36], the dominant color region map $DCRM(i, j)$ as follows:

$$DCRM(i,j) = \begin{cases} 0 & D(i,j) > D_{th} \\ 1 & D(i,j) \le D_{th} \end{cases} \tag{5}$$

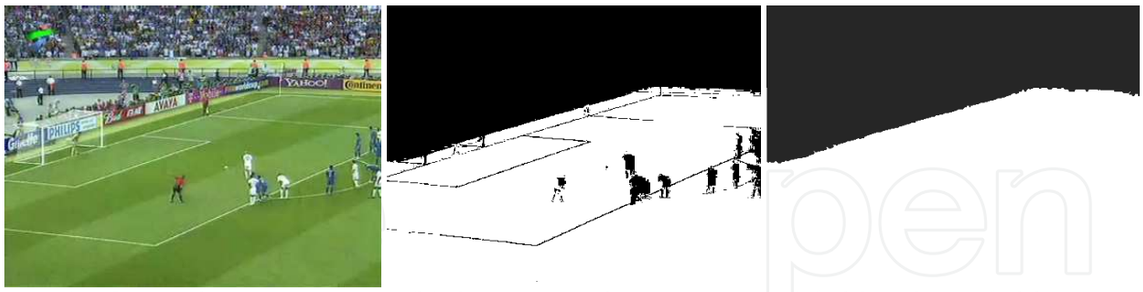
where D_{th} is learned from several soccer video clips, $DCRM(i, j)$=1 indicates that the pixel (*i*, *j*) belongs to the field region. The dominant color ratio (*DCR*) of a frame is obtained as follows:

$$DCR = \frac{\sum_{i=1}^{H}\sum_{j=1}^{W} DCRM(i,j)}{H \times W} \tag{6}$$

where *H* and *W* are the height and width of a frame. Fig.2 shows the corresponding dominant color region extraction results. The dominant color region of Fig.2(a) is shown in Fig.2 (b) and the field region is extracted as shown in Fig. 2 (c). The dominant color distribution is represented by the dominant color projection vectors of *DCRM*.

Based on *DCRM*, both global or grid based dominant color distributions can be utilized. In [56], the *DCRM* is parsed into 8 equal-sized regions in both vertical and horizontal direction. By calculating the dominant color pixel ratio of each region, a 16-bin dominant color distribution vector is constructed. In addition to the dominant color distribution, a 255

dimensional block wise color moment generated from 5-by-5 grids of the images is also utilized.



(a) Original image (b) dominant color region (c) field and non-field region

Fig. 2. Dominant color region, field region and field lines detection results.

2.2 Texture feature

Texture features are insensitive to lighting, rotation, and other conditions. Two kinds of texture features are utilized for coarse shot classification in [56]. The first is hierarchical wavelet packet texture descriptor (HWVP) [47]. The second is 24 dimensional histogram of oriented gradient (HOG) [46]. HWVP descriptor is a 210 dimensional feature which is extracted under local partitioning pattern Local5 (the image is partitioned into 2×2 grids and a centralized grid), by setting the wavelet packet basis to be db2, with hierarchical wavelet packet transform level 2.

2.3 Title text detection, localization, and tracking

Title texts provide valuable information for event detection in soccer video content analysis. Usually the overlaid text about goal, foul, yellow/red card are followed the corresponding highlight events. In Fig.3, overlaid texts (including bulletin texts and title texts) and replays of a soccer video are plotted. There are 20 replays and 7 overlaid texts. The first text is a bulletin text that shows the players' name of the two teams. The second text shows the initial score of two teams. The 3rd and the 6th texts show the names of the players who got goals. The 4th and 7th texts show the updated scores just after a goal. Moreover, the 5th text shows the score of two teams. From Fig.3, the co-occurrences of the title texts and replays always indicate the appearing of highlights. So, it is reasonable to fuse the title text detection results to improve the highlight detection performance. In [56], the detected title text information is utilized for improving highlight detection performances. Thus from the production knowledge overlaid text information is one of the effective clues for high-level semantics inference.

There are four types of texts in the sport video sequences: 1) long term texts, such as icons of the TV channels, the score-boards which appear at the four corners of video frame, as shown in Fig.4(a) and (b) respectively. This type of texts exists in a fixed location in a frame with long duration. 2) Title texts, e.g. showing the score of two teams after getting a goal, as shown in Fig.4(b). 3) Scene texts, such as the texts appearing in signboard and cloth. 4) Bulletin texts, e.g. showing the name list of a team. This type of text usually appears at the beginning of a soccer video. Except utilizing the text detected in sports video, web-casting

text information is incorporated with audio-visual features to improve highlights detection performances [28,29].

In soccer video the title texts usually appeared in the bottom-center of the image. The icon, scoreboard and time tables are appeared in the top-left and top-right parts of an image respectively. This type text always appeared during the whole video sequence, for example, the extracted local text lines of Fig. 4(a), as shown in Fig.4(c) are appeared in the whole video. However, the title text as appeared in Fig.4(b), as shown in Fig. 4(d) usually existed in a limited time range. The title text is usually provides more information on its corresponding highlight event inference than a long term text. They are purposively added during soccer video editing, and they are aimed at providing the audience some indicative information about the video content, such as a card or a goal [56], [62].

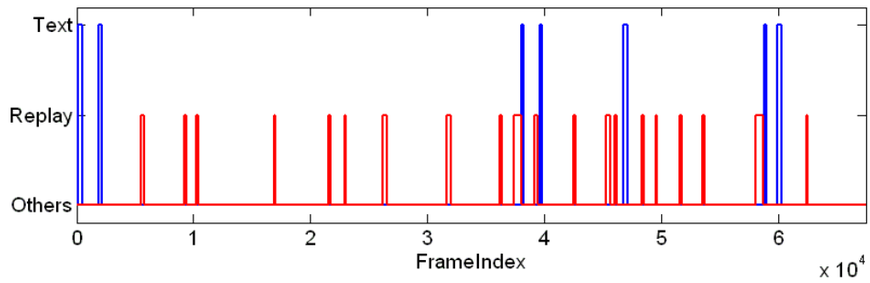


Fig. 3. Overlaid texts (including bulletin texts and title texts) and replays of a soccer video.

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

Fig. 4. The types of texts appeared in soccer video frames.

2.3.1 Related works on video text detection and tracking

The corresponding text detection approaches can be classified into following 3 types. The first is the connection characteristics of video texts [63], [64]. The text detection methods based on this characteristic assume that text regions have uniform colors and satisfy certain constraints on size, shape, and spatial layout. The second is the texture alike characteristic of the text regions [65, 33, 66, 67]. The text detection methods based on texture information usually assume that the text regions have special texture patterns. And the third is the edge density information [34], [68]. These methods make full use of the fact that the edge densities of background are comparatively sparser than those of the text regions [69], [70]. Usually the corner point number in the text region is larger than that in the background regions.

Video text detection and localization can be carried out both in pixel and compressed domains [71,34,68,65,72]. In order to eliminate false detections, the edge [69], texture [33], and shape information [57] are often utilized in text verifications. In addition, the available redundant temporal information is often used in candidate text region verification and falsely detected text region elimination [71,34,68,65,72,73]. Lyu et al. proposed a multi-resolution based text detection method [34]. Firstly, original edge map is generated for each of target video frames. Multi-resolution text maps of a target frame are generated by down-sampling the original text map. Then text detection, verification and localization are carried out on multi-resolution text maps. Finally, the text detection texts in various resolutions are integrated.

2.3.2 Text detection and tracking

In this chapter, we utilize the corresponding text detection and tracking approaches [56]. The block diagram of the title text detection, localization and tracking is shown in Fig.5. Text line number, spatial location, and temporal duration constraints are utilized during title text detection, localization and tracking. Lyu et al' method [34] is used to detect and localize title texts for a given frame.

It is not necessary to carry out text detection for each of the video frames. In our observations, we find that title texts usually exist for about 5 seconds in soccer video. Detecting text with an interval of one or half second is enough, e.g. detecting text in intra-frames of sport video. Only the texts appeared at the bottom-center of the images and with sufficient duration are determined as candidate title texts. The starting and ending frames of each title text are determined by text line matching and tracking [33].

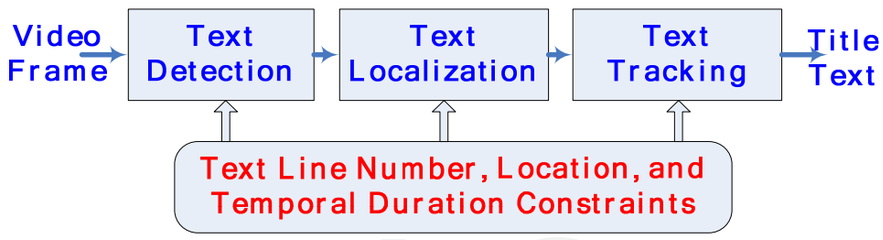


Fig. 5. Title text detection, localization and tracking with text line number, localization and temporal duration constraints.

2.4 Motion feature representation

Motion information is also very important for sport video content analysis [14,15,17]. Typically the camera men operate cameras, by fast track, or zoom in to provide audiences with clearer views of the games. Based on the camera motion (i.e. global motion) patterns and domain related knowledge, high level semantics can be inferred.

In [18], global views of soccer video are further refined into the following three types: stationary, zoom and track in terms of camera motion information using a set of empirical rules with respect to domain and production knowledge. The key-frames of a shot with stationary by means of average motion intensity and average motion intensities of global motion.

Global motions in a video sequence are caused by camera motion, which can be modeled by parametric transforms [30, 32]. The process of estimating the transform parameters is called global motion estimation. The widely used global motion model is perspective model with 8 parameters, which is expressed as follows

$$\begin{cases} x' = \dfrac{m_0 x + m_1 y + m_2}{m_6 x + m_7 y + 1} \\ y' = \dfrac{m_3 x + m_4 y + m_5}{m_6 x + m_7 y + 1} \end{cases} \tag{7}$$

where (x, y) and (x', y') are the coordinates in the current and the reference image respectively, with the set of parameters $\mathbf{m} = [m_0, \cdots, m_7]$ denoting the global motion parameters to be estimated.

Average motion intensities of global motion and local motion are utilized for coarse semantic refinement [56]. The average global motion intensity ($AGMV$) is calculated as follows:

$$AGMV = \frac{1}{M} \sum_{j=1}^{M} \sqrt{GMVx_j{}^2 + GMVy_j{}^2} \tag{8}$$

where $(GMVx_j, GMVy_j)$ is the global motion vector of the block at (x_j, y_j), M is the total block number. The global motion vector $(GMVx_t, GMVy_t)$ at the coordinates (x_t, y_t) is determined as follows

$$\begin{cases} GMVx_t = x'_t - x_t \\ GMVy_t = y'_t - y_t \end{cases} \tag{9}$$

where (x'_t, y'_t) are the warped coordinates in the reference frame by the global motion parameters from the coordinate (x_t, y_t).

The local motion information is represented by average motion intensity (AMV) which is expressed as follows

$$AMV = \frac{1}{M} \sum_{j=1}^{M} \sqrt{MVx_j{}^2 + MVy_j{}^2} \tag{10}$$

where (MVx_j, MVy_j) is the motion vector (MV) of the block with its coordinates (x_j, y_j) and j is the block index.

2.5 Scene change detection and logo detection

Parsing the sequential video sequences into shots is helpful for video content analysis. This is often called scene change detection or shot boundary detection. Thee sport video sequences are different with other video sequences, e.g. the highlights are often replayed. Usually, logos are utilized to connect the live broadcasted clips with the replayed segments. Thus, scene changes and logos are the basis of sport video semantics detection and semantic based sport video content browsing. Scene change detection and logo detection approach

consists of the following three steps [3,56]: (1) logo template detection based on the fact that different starting or ending logo transitions, (2) logo transitions detection in the video program using the logo template based on the probability models of pixel-wise intensity-based mean-square difference and color histogram-based mean-square difference, (3) the replay segments identification.

3. Mid-level semantic classification

In this section, the corresponding middle level semantics for sport video are defined and detected. Related work on middle level semantic classification is reviewed.

3.1 Related work on mid-level semantic classification

Xu *et al*. classified each soccer video shot into one of the following 3 views : global, zoom-in and close-up [16]. From the view labels, soccer video sequences are further parsed into play and break. Duan *et al*. classified video shots into predefined mid-level semantics [18]. Based on which, soccer video is coarsely parsed into two type in-play and out-of-play [14]. In [43], global motion and visual features are fused to carry out shot categorization for basketball video. Tan *et al*. also segmented a basketball sequence into wide angle, close-up, fast break and possible shoot at the basket [61]. Motion and dominant color information of a video shot are fused by multi-layer hidden Markov models (HMM) to determine its category [2]. In [37], each shot of a baseball video is classified into predefined semantic scenes by fusing features extracted from visual content of the image, object level feature representations, and camera motion.

(a) Global view (b) Medium view

(c) Close-up (d) Audience

Fig. 6. Four coarse views for soccer video.

3.2 Coarse mid-level semantic classification

Fig.6 shows four coarse mid-level semantics for soccer video. They are global view, medium view, close-up and audience respectively.

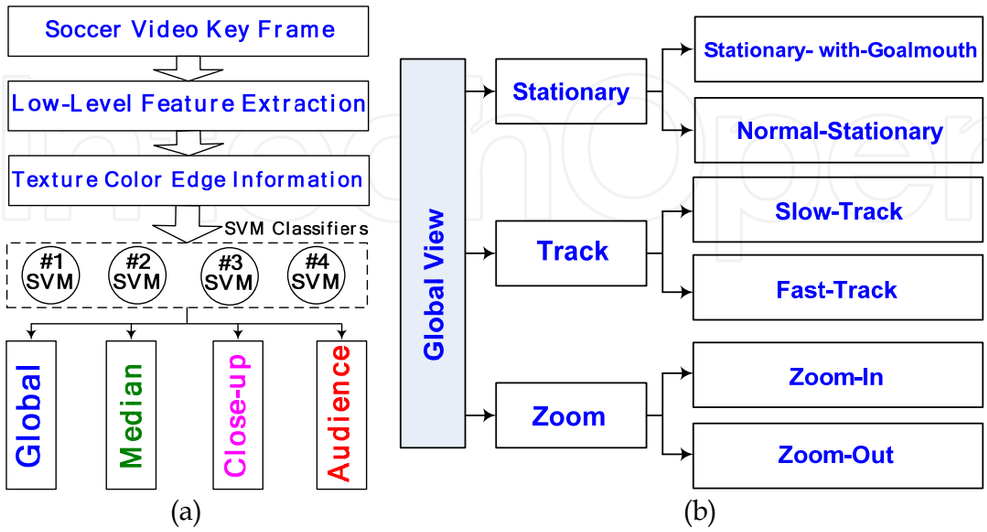


Fig. 7. Coarse to fine soccer video shot categorization. (a) Coarse soccer video shot classification using four one-versus-all SVM classifiers. Replay is refined into replay-global, replay-median and replay-close-up using the first three SVM classifiers. (b) Global view refinement based on camera motion information and field line detection results.

In coarse mid-level semantic categorization, soccer video shots are parsed into four coarse views of global, median, close-up, and audience using four one-versus-all SVM classifiers as shown in Fig.7 (a). The kernels of SVM classifiers are RBF (radius-basis-function). The input of the four SVM classifiers is a 505 dimensional low-level feature, including 255d color moment, 16d dominant color distribution, 24d HOG feature, and 210d HWVP feature.

3.3 Refinement for coarse mid-level semantics

As median, audience, and close-up views are related to the details, which do not need further refinement. Moreover, the audience shots are rarely replayed, the replay into three types: replay-global, replay-median and replay-close-up using the trained SVM classifiers for coarse semantics classification [56]. The block diagram of global view refinement is shown in Fig.7 (b) respectively.

Global view is further refined into one of the following three types: stationary, zoom and track with respect to camera motion information using a set of empirical rules. A shot is determined as with stationary if AMV<0.5, otherwise non-stationary. The non-stationary shot is further refined into track if m_0=m_5 =1, otherwise zoom. A zoom is a zoom-in if m_0=m_5 >1 and a zoom-out if m_0=m_5 <1. The track is a slow-track if $AGMV \le 2$, otherwise a fast-track.

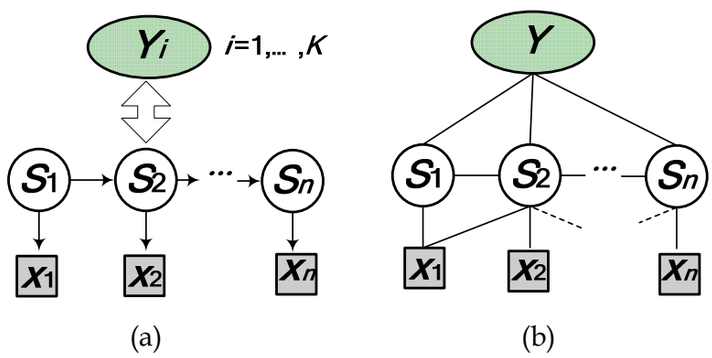
A stationary shot is further refined into stationary-with-goal-post and normal stationary according to the field lines detection results. When the valid field line number is larger than four, then the frame is a stationary with goal-post (SG), otherwise a normal stationary [56].

After mid-level semantic classification and refinement, each segment of a soccer video or event clip is classified into one of 13 mid-level semantics: logo, audience, median, close-up, replay-global, replay-median, replay-close-up, fast-track, slow-track, zoom-in, zoom-out, normal stationary, and stationary-with-goalmouth.

4. High-level semantic detection

In this section, the high level sport video semantics detection approaches are illustrated in details. Statistical learning models such as hidden Markov model (HMM) as shown in Fig.8(a), enhanced hidden Markov model (EHMM), and hidden conditional random field (HCRF) as shown in Fig.8 (b), based soccer video event detection approaches are described. Correspondingly, soccer video highlight detection results are given.

The main content of this section is as follows. Firstly the related works on soccer video highlight detection are briefly overviewed in Section 4.1. Secondly, event clip segmentation in Section 4.2. Thirdly, HMM, EHMM and HCRF based soccer video event detection approaches are provided in Section 4.3 and Section 4.4 respectively. And finally, highlight detection performances are evaluated in Section 4.5.



(a) (b)

Fig. 8. HMM and HCRF models. (a) HMM and (b) HCRF.

4.1 Related work on soccer video highlight detection

As shown in Fig.1, one type of highlights detection approaches are carried out by mapping from low-level audio-visual features [1], [2], [14], [23], mid-level semantics, coarse events [23], [25], scenes with overlaid text-lines [31] and replays [3]-[5], [31]. Another type of highlight detection approaches are carried out by statistical learning based models. The learning based approaches have been proved to be effective in fusing multi-modal features to improve highlights detection performance [6], [11], [15]-[21], [23]-[27], [35]-[37], [39]-[42], [44], [45].

Xie *et al.* proposed multi-HMMs based method to improve the performance of play-break detection [6]. The HMMs are different structures and with the same input. Dynamic programming is utilized to fuse the outputs of multi-HMMs to determine the shot type (play or break). In [26], soccer video is segmented into sequential event clips: forward pass, shot on goal, placed kick, turnover, counter-attack, and kick-off using a set of domain rules and sport video production knowledge. The placed kicks are further refined into corner kick, free kick and penalty according to the distribution of players' position.

Wang *et al*. proposed conditional random fields (CRF) based soccer video event detection method [11]. Firstly, mid-level semantics are determined by mapping from low-level audio-visual features. Then, highlights are detected by fusing the mid-level semantic using CRF [11]. In [27], dynamic Bayesian networks (DBN) are utilized to model goal, corner kick, penalty, and foul events. Low-level audio-visual features are mapped into mid-level nodes of Bayesian networks directly. Bayesian networks model the dependency of the observations of each shot for event type inference. The shot-based event recognition results are fused by DBN to accumulate evidence in the temporal domain [27].

HCRF is utilized to model highlight events in golf and bowling videos [60]. Different from the existing event detection approaches, the transformed low-level features are utilized to perform event detection. Independent component analysis (ICA) is utilized to determine two main components of the input low-level feature. The HCRF is utilized to fuse the ICA components for event type determination.

In most existing works, events are determined from shot level. It is well known that, a single shot separated from its context does poorly in conveying semantics [22], [56]. That is to say the contextual information among the shots of an event clip is not fully disclosed in event inference [11], [27]. Grouping the sequential shots with the same semantics into unified event clips and then recognizing their event types is an optimal way in event detection [56], [21].

4.2 Event clip segmentation

Let *VS* denote a video sequence. Assuming that it consists of *K* sequential events, the whole video sequence can modeled as follows

$$VS = (E_1, \cdots, E_K) \tag{11}$$

where $E_t (t = 1, \cdots, K)$ belongs to one of the predefined events. Each event clip E_t is composed of several sequential mid-level semantics, which is expressed as

$$E_t = (M_t^1, \cdots, M_t^{N(t)}) \tag{12}$$

where M_t^q ($q = 1, \cdots, N(t)$) corresponds to the predefined mid-level semantics, which can be represented by key-frames of video shots. $N(t)$ is the total number of mid-level semantics of the *t*-th events. From Eq.(12) and Eq.(11), we have

$$VS = \left(\underbrace{M_1^1, \cdots, M_1^{N(1)}}_{E_1}, \cdots, \underbrace{M_t^1, \cdots, M_t^{N(t)}}_{E_t}, \cdots, \underbrace{M_K^1, \cdots, M_K^{N(K)}}_{E_K} \right) \tag{13}$$

We aimed at segmenting *VS* into event clips by finding event boundaries according to the consistence of semantics. According to domain rules and production knowledge of soccer video, the starting and ending frames of an event clip are the frames at the boundaries where non-global views (including median, close-up, audience, replay and logo) transition to global views [56].

Fig. 9 shows a diagram of event clip (EC) segmentation. The event clips #1 - #4 are composed of global views and several close-up or median views. The event clip #4 is composed of the following mid-level semantics: global, close-up, median, close-up, logo, replays, and logo.

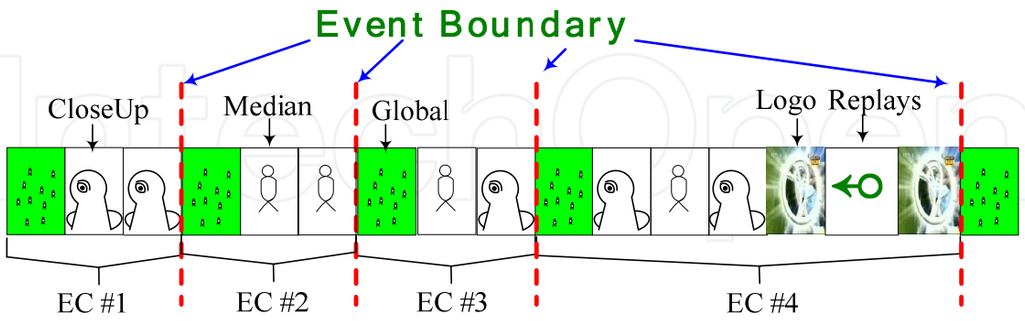


Fig. 9. An example of event boundary determination.

4.2.1 Observations of an event clip

Assuming that an event clip is composed of n mid-level semantics, x_j ($j = 1,\cdots,n$), the temporal transitions of the mid-level semantics of an event clip can be viewed as observations of statistical learning models. Moreover, in [56], the overall feature extracted from each event clip is combined with the temporal observations for event type inference.

The normal observations of an event clip are the temporal transitions of mid-level semantics (i.e. $\mathbf{x} = (x_1,\cdots,x_n)$) as shown in Fig.10. In [56], the enhanced observations $\mathbf{x} = (x_1,\cdots,x_n,\cdots,x_{n+k})$ consist of the normal observations and the overall features are utilized to accumulate the semantics for soccer video highlights detection.

The enhanced observation is composed of k overall features ($k \le 3$). They correspond to the title text (TLT), long time break (LTB), and replay (REP) information respectively. All of them are binary. TLT=0 means that there is no title text in event clip. REP =0 denotes that there is no replay in this event clip. LTB=0 represents that the non-global view segment number is very small.

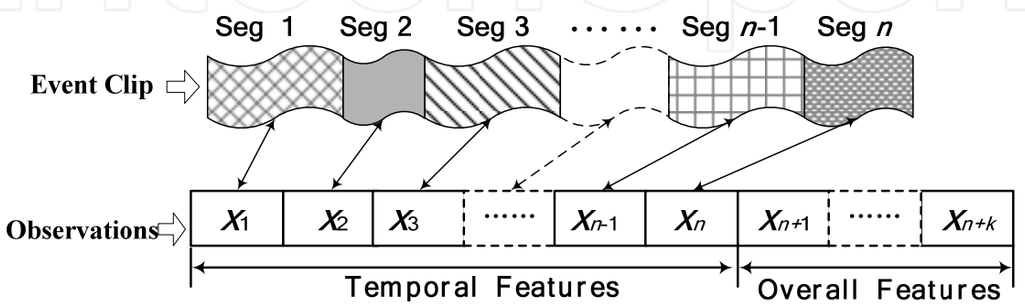


Fig. 10. The temporal and overall features of an event clip. k is the overall feature number.

4.3 HMM based event detection approach

HMM models the state sequence as being Markov, and each observation being independent of all others given the corresponding state [52]. HMM models the joint distribution under two basic independence assumptions: Markov property and independent property.

In this section the hidden Markov and enhanced Markov models based event detection approaches are illustrated in details. The HMM based approach is carried out event inference using the normal observations of an event clip. While, the EHMM based approach carries out event inference using the enhanced observations of an event clip.

4.3.1 Overview of HMM

HMM is a generative model. It defines a joint probability distribution for the observations and their corresponding labels. HMM models a sequence of observations $\mathbf{x} = (x_1, \cdots, x_n)$ with the corresponding label *y* under the assumption that there is an underlying sequence of state $\mathbf{s} = (s_1, \cdots, s_n)$ drawn from a finite state set. HMM carries out inference under two basic independence assumptions [43]. The first assumption is that each state s_j depends only on its immediate predecessor $s_{j\text{-}1}$, and independent of its previous states (i.e. $s_1, \cdots, s_{j-2}$). The second assumption is that each x_j depends only on the corresponding state s_j. Let $\lambda = (A, B, \pi)$ denote the model parameters. A denotes the state transition probability distribution $A = \{a_{jk}\}$, where a_{jk} is the state transition probability from state j to state k ($a_{jk} \geq 0$). B denotes the observation probability distribution in state j with its observations x_k, $B = \{b_j(k)\}$. $\pi = \{\pi_j\}$ is the initial state distribution. Thus the joint probability of a state sequence **s** and an observation sequence x under a model λ can be modeled as follows

$$P(\mathrm{x} \mid \mathrm{s}, \lambda) = \prod_{j=1}^{n} P(x_j \mid s_j, \lambda) \tag{14}$$

By assuming that observations are statistical independent, then we have

$$P(\mathrm{x} \mid \mathrm{s}, \lambda) = b_{s_1}(x_1) \times b_{s_2}(x_2) \times \cdots \times b_{s_n}(x_n) \tag{15}$$

4.3.2 The training of HMM

Let $D = \{(\mathrm{x}^{(k)}, y^{(k)})\}_{k=1}^{N}$ denote the training data. *N* is the total number of training event clips. For the *k*-th training clip, its input is $\mathrm{x}^{(k)} = \{x_1^{(k)}, x_2^{(k)}, \cdots, x_T^{(k)}\}$ and its label is $y^{(k)}$. $x_m^{(k)}$ corresponds to the *m*-th mid-level semantics in the *k*-th training clip, and $y^{(k)}$ is its event label.

The flowchart of the training of HMMs is as follows. Firstly, the mid-level semantics of each event clip are determined from the extracted low-level features. Then, the Expectation-maximization (EM) is utilized to estimate the model parameters $\lambda_i = (A_i, B_i, \pi_i)$. The EM algorithm consists of E-steps and M-steps [43]. The E-step computes the forward and backward probability for the given model, and the M-step re-estimates the model parameters. Given an initial model λ, the EM algorithm can find a model λ^t at *t*-th iteration

such that its performance is better than the (*t*-1)-th iteration. That is to say for the *K* training samples, the EM algorithm makes the following equation hold.

$$\sum_{k} P(\mathbf{x}^{(k)} \left| \lambda^{t}) \geq \sum_{k} P(\mathbf{x}^{(k)} \right| \lambda^{t-1}) \tag{16}$$

4.3.3 The classification of HMM

In event recognition, the type of the event clip with observations **x** is determined as follow

$$y = \arg\max_{i} P(\mathbf{x} \left| \lambda_i) = \arg\max_{i} \sum_{s \in \mathbf{S}} P(\mathbf{x} | \mathbf{S}, \lambda_i) \times P(\mathbf{S} \right| \lambda_i), \tag{17}$$

The probability indicates how well the model λ_i matches the given observations **x**. The probability of the hidden state sequence S can be expressed as

$$P(\mathrm{s} \mid \lambda) = \pi_{s_1} a_{s_1 s_2} \times a_{s_2 s_3} \times \cdots \times a_{s_{n-1} s_n} \tag{18}$$

Thus, Eq.(17) can be rewritten as

$$y = \arg\max_{i} \sum_{s_1, \cdots, s_n} \pi^i_{s_1} b^i_{s_1}(x_1) a^i_{s_1 s_2} \times b^i_{s_2}(x_2) a^i_{s_2 s_3} \times \cdots \times b^i_{s_n}(x_n) a^i_{s_{n-1} s_n}, \tag{19}$$

4.4 HCRF based event detection

Hidden conditional random fields (HCRF) are discriminative models that generalize the hidden Markov models (HMM) and the conditional random fields (CRF) [52]. HCRF models the state sequence as being conditionally Markov given the observations. Unlike HMM, HCRF is also capable of modeling long range dependencies of the observation.

4.4.1 Overview of HCRF

HCRF uses intermediate hidden variables to model the latent structure of the observations as shown in Fig.8 (b). A HCRF models the conditional probability of a label y given the observations $\mathrm{x} = \{x_1, \cdots, x_n\}$ with latent states $\mathrm{s} = \{s_1, \cdots, s_n\}$ as follows:

$$p(y \mid \mathrm{x}; \lambda) = \frac{1}{Z(\mathrm{x}; \lambda)} \sum_{\mathrm{s}} p(y, \mathrm{s} \mid \mathrm{x}; \lambda) \tag{20}$$

where $p(y, \mathrm{s} \mid \mathrm{x}; \lambda)$ is a conditional probability, given the labels y, observations x, and hidden states s under the HCRF model parameters λ.

$$p(y, \mathrm{s} \mid \mathrm{x}; \lambda) = \exp\{\Psi(y, \mathrm{s}, \mathrm{x}; \lambda)\} \tag{21}$$

where $\Psi(y, \mathrm{s}, \mathrm{x}; \lambda)$ is a potential function parameterized by the model parameters λ. $Z(\mathrm{x}; \lambda)$ is a normalization factor. It ensures that the model be a properly normalized probability over all labels. It is defined as follows

$$Z(\mathrm{x};\lambda)=\sum_{y'}\sum_{\mathrm{s}}\exp\{\Psi(y',\mathrm{s},\mathrm{x};\lambda)\} \tag{22}$$

where y′ is a possible label for the observations x.

4.4.2 The training of HCRF

The parameters of HCRF are trained on the training data $D=\{(\mathrm{x}_i,y_i)\}_{i=1}^{N}$, where each observation $\mathrm{x}_i=\{x_{i,1},x_{i,2},\cdots,x_{i,n}\}$ is a sequence of inputs, and each y_i is the label index of the input sequence x_i. The parameters estimations in HCRF is carried out by using the following objective function

$$L(\lambda)=\sum_{t\in\{1,\cdots,N\}}\log p(y_t\mid \mathrm{x}_t;\lambda)-\frac{1}{2\sigma^2}\|\lambda\|^2 \tag{23}$$

The last term $\frac{1}{2\sigma^2}\|\lambda\|^2$ is used for regularization which can reduce the over-fitting of the objective function in optimization. It is the log of a Gaussian prior with variance σ^2 [48, 49]. The limited memory BFGS can be utilized to find the optimal parameters λ^* as follows [50]-[52].

$$\lambda^*=\arg\max_{\lambda}L(\lambda) \tag{24}$$

4.4.3 The classification of HCRF

Given a test event clip with its observations $\mathrm{x}=\{x_1,\cdots,x_n\}$ and trained parameters λ of the HCRFs, the label *e** of this event clip is recognized as follows

$$e^*=\arg\max_{e}P(e\mid \mathrm{x},\lambda_e) \tag{25}$$

where λ_e is the parameter vector of the event *e*. In this chapter, we have *e*∈{Goal, Shoot, Normal Kick, Foul, Placed Kick}.

4.5 Highlight detection performances evaluation

In this section, high-level semantics detection performances of HMM, EHMM, and HCRF based approaches are evaluated. The training set for the models HMM, EHMM and HCRF is the same which consists of *N* manually labeled examples $(\mathbf{x}^i,y^i)$ $(i=1,\cdots,N)$ where each y^i∈{Goal, Shoot, Normal Kick, Foul, Placed Kick}is the event label, and each observations are $\mathrm{x}^i=\{x_1^i,\cdots,x_n^i\}$. The component x_k^i corresponds to the observation of the *k*-th segment ($k\in\{1,\cdots,n\}$) of the *i*-th training event clip. 12 soccer sequences and some highlight event clips which are downloaded from Internet are utilized to train the model parameters of HMM, EHMM and HCRF.

The training steps are as follows: 1) manually label the event clips for the test soccer video sequences. 2) manually label the mid-level semantics. 3) manually label the event boundaries; 4) get the corresponding mid-level semantics for each event clip.

The observations of HMM and HCRF are the temporal transitions of mid-level semantics. The observations of EHMM consist of two parts. The first part is temporal transitions of mid-level semantics which is identical to the observations of HCRF. The second is enhanced observations, including the title texts, long term breaks, and replay information of an event clip [56]. Three overall features are utilized in EHMM. The EHMM and HCRF based event detection approaches are on the basis of event clips and the observation of EHMM and HCRF are mid-level semantics of an event clip.

In order to show the performances of model based event detection approach, in this Section the HMM, EHMM and HCRF based soccer video highlights detection performances are evaluated on seven soccer video sequences. These video sequences are captured from a variety of sources. The total duration of the test video sequences is about 625 minutes. Totally, there are 26 goals, 137 placed kicks, 85 fouls, and 109 shoots. Recall *NR*, precision *NP* and F-measure *F* are used to evaluate objective event detection performances, which are defined as follows:

$$NR = \frac{NC}{NC + NM} \times 100\% \tag{26}$$

$$NP = \frac{NC}{NC + NF} \times 100\% \tag{27}$$

$$F = \frac{2 \times NR \times NP}{NR + NP}\% \tag{28}$$

where *NC*, *NM*, and *NF* denote the correctly, missed, and falsely detected event numbers. Moreover, confusion matrix is utilized to show the discrimination of HCRF based event detection approach for the five events.

The recall, precision and F-measure values of the four highlights of HMM, EHMM and HCRF are shown in TABLE I. The average recall values of the four highlight events are79.55%, 89.08%, and 86.55%. The average precision values of the four highlight events are76.96%, 87.60%, and 88.03%. The average F-measure values of the highlights are 78.24%, 88.33% and 87.29% respectively.

	Placed Kick			Foul			Shoot			Goal		
method	NC	NM	NF	NC	NM	NF	NC	NM	NF	NC	NM	NF
HMM	113	24	18	69	16	25	81	28	28	21	5	14
EHMM	130	7	2	76	9	11	88	21	22	24	2	10
HCRF	120	17	3	73	12	13	90	19	14	26	0	12

Table 1. Comparisons of highlights detection performances of HMM, EHMM and HCRF.

EHMM and HCRF based highlight detection approaches outperform that of HMM based approaches. EHMM is a little better than the HCRF based approach for the events: placed kick and foul. The main reasons are due to the following two aspects: 1) overall features of an event clip are utilized in EHMM. The overall features served as global features which are helpful for highlight event discrimination. 2) The overall features compensate the interior of HMM in modeling the dependence of long term observations and ease the two basic

assumptions of HMM. However, HCRF achieves highest performances for detecting shoot and goal events.

5. Semantic based sport video browsing

In this section, a semantic based video browsing framework is proposed. Users can view the video content freely like reading books. The semantic based video content browsing approach is organized as the table-of-content (ToC) of a book. Let's give a brief overlook of the ToC of a book, before illustrating the proposed video browsing method. Fig.11 shows the ToC structure of a book which can be departed into following seven layers: (1) book title (BT); (2) chapter (CH); (3) subchapter (SC); (4) paragraph (PH); (5) page (PG); (6) sentence (ST); and (7) words (WD). With respect to the ToC, readers can know its content very well. They can go straight to the interested content by skipping the irrelevant parts. In order to let readers know the overall content of book, the **preface**, **introduction** and **summary** of a book give the brief illustrations of the book, a chapter and a subchapter.

If sport video content is organized as the ToC of a book, it will be convenient for us to browsing its content. In this chapter, a novel book style based semantic video browsing approach is expressed. TABLE II gives the correspondences of ToC of book and soccer video. In the first layer the book title corresponds to the video category, such as baseball and soccer video. In the second layer, the chapter of a book is similar to a set of events. In the third layer, the sub-chapter corresponds to the detailed information of an event. In the fourth layer, the paragraph of a book is corresponding to the shot of an event. In the fifth layer, the page of a book is similar to the frame of a video. In the sixth layer, the sentence of a book is corresponding to the object in a video sequence. And in the seventh layer, the word of a book corresponds to the pixel. Fig.12 shows the ToC of soccer videos. Soccer video is parsed into following seven events: normal kick, foul, free kick (FK), penalty, corner kick, shoot and goal. The free kick, penalty and corner kick are refined from placed kick using the distributions of players' positions [7]. The fifth layer of the video provides the original video sequences of event.

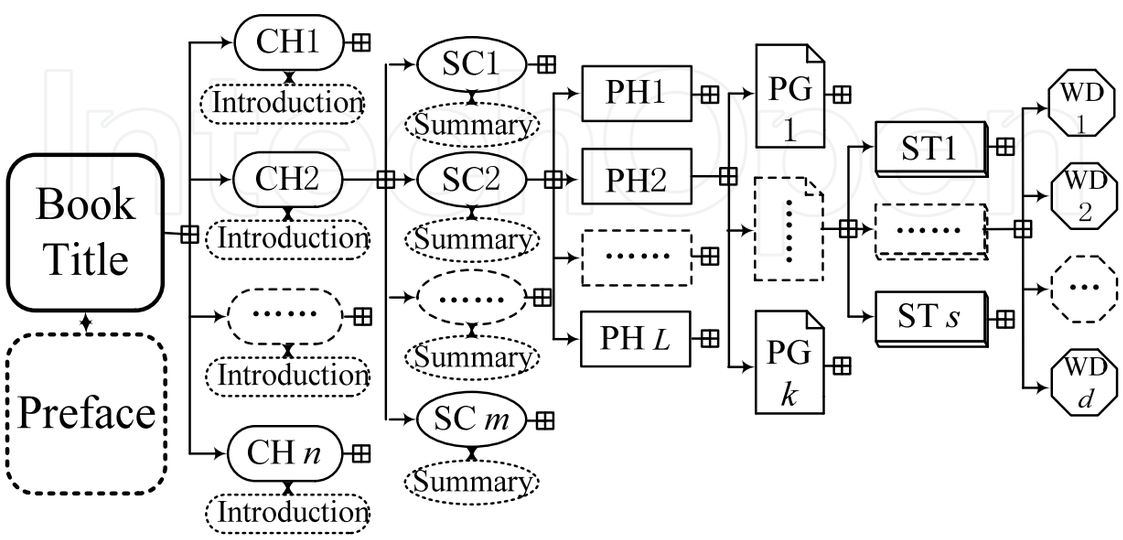


Fig. 11. Table-of-Content of a Book.

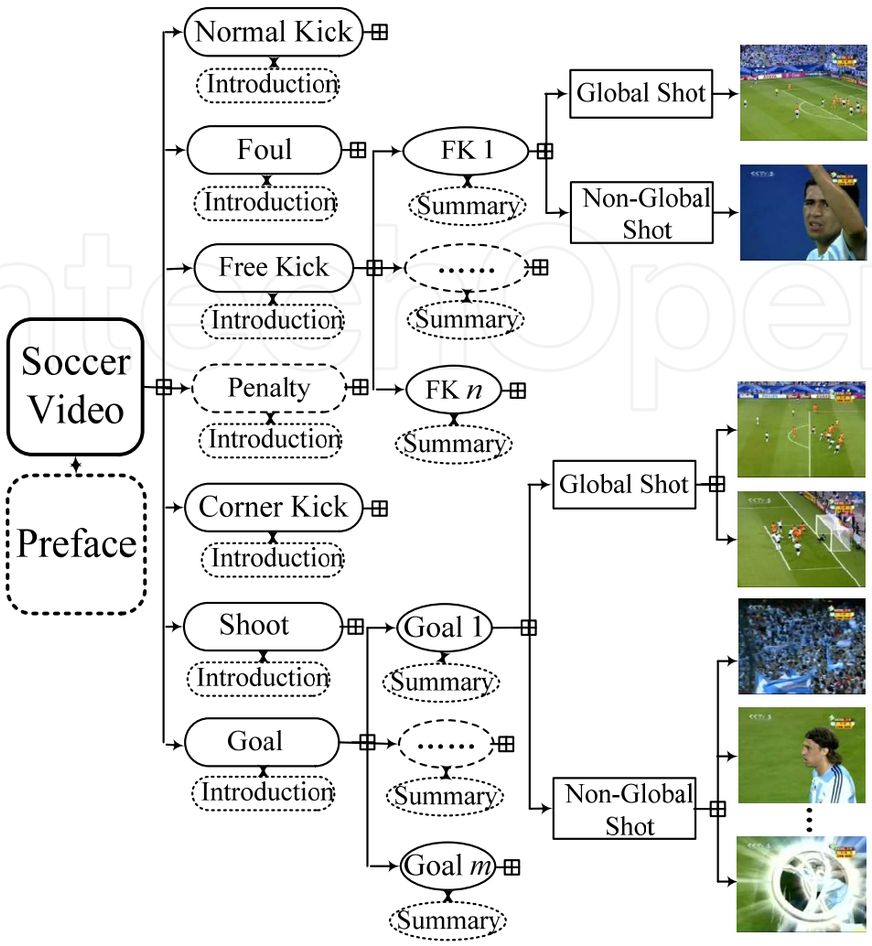


Fig. 12. ToC of soccer video.

The target of us is how to carry out ToC based sport video summarization. According to the correspondence of ToC of Book and sport video, a four layer summarization framework is utilized in this chapter for soccer video summarization. In the fourth layer, the abstraction is carried out for the semantic shots by extracting several key-frames. The third layer abstraction (i.e. **summary**) of a video is the shortened video frames of a specified story unit or event. The summary of the ToC of a video sequences in the third layer is the combination of the summarizations extracted in the fourth layer for the semantic shots. In the sports video, there are also two kinds of semantic shots: global shots and non-global shots. The second layer abstraction (i.e. **introduction**) of a video is the skimmed video frames of story units or events of a certain class. The first layer abstraction (i.e. **preface**) of a video is providing the corresponding abstractions of the video sequences.

Fig.13 shows an example of the proposed sematic based soccer video browsing approach for a goal event. In the fourth layer, it is the continuous video clips of this event. The third layer shows the extracted 11 key-fames of this event clip, which corresponds to the summary of this event. The second layer shows two representative key-frames, which corresponds to the introduction of this event. The first layer is the most representative frames of this event, namely the preface of this event.

Layer	Book	Sports Video
1	Book Title	Soccer video
2	Chapter	A set of Events
3	Sub-Chapter	An Event
4	Paragraph	Mid-level Semantics
5	Page	Frame
6	Sentence	Objects
7	Word	Pixel

Table 2. The Correspondence of ToC structure of different types of video sequences.

(a) The third layer abstraction of an event (summary of ToC)

(b) The second layer abstraction of an event (introduction of ToC)

(c) The first layer abstraction of an event (preface of ToC)

Fig. 13. Video ToC for a soccer event.

6. Conclusion

In this chapter, semantic based sport video browsing is introduced. Soccer video high-level semantics detection approaches using hidden Markov models, enhanced Markov models, and hidden conditional fields are expressed in detail and their performances are evaluated. From the detected highlights, a semantic based soccer video browsing approach is proposed. The proposed semantic based soccer video browsing approach carries out video content browsing using a book-like structure.

7. Acknowledgement

This work is supported in part by National Natural Science Foundations of China (NSFC) No.60903121, No.61173109, and Foundation of Microsoft Research Asia.

8. References

[1] B. Li, J. Errico, H. Pan, and M. Sezan, "Bridging the semantic gap in sports video retrieval and summarization," *J. Vis. Commun. Image R.* vol.17, pp.393-424, 2004.

[2] G. Xu, Y. Ma, H. Zhang, and S. Yang, "An HMM-based framework for video semantic analysis," *IEEE Trans. Circuits and Systems for Video Technology*, vol.15, no.11, Nov. 2005, pp.1422-1433.

[3] H. Pan, B. Li, and M. Sezan, "Automatic detection of replay segments in broadcast sports programs by detecting of logos in scene transitions," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, vol.4, pp. 3385-3388, Orlando, FL, May 2002.

[4] Z. Zhao, S. Jiang, Q. Huang, and G. Zhu, "Highlight summarization in sports video based on replay detection," in *Proc. Int. Conf. Mulmedia and Expo.*, pp. 1613-1616, Toronto, Ontario, Canada, July 2006.

[5] H. Pan, P. Beek, and M. Sezan, "Detection of slow-motion replay segments in sports video for highlights generation," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, vol.3, pp.1649-1652, Salt Lake City, USA, May, 2001.

[6] L. Xie, S. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with hidden Markov models", in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*,2002, pp. 4096-4099.

[7] A. Ekin, and A. Tekalp, "Generic play-break event detection for summarization and hierarchical sports video analysis," in *Proc. Int. Conf. Mulmedia and Expo*, vol.1, 2003, pp. 169-172.

[8] D. W. Tjondronegoro, Y. Chen, and B. Pham, "Classification of self-consumable highlights for soccer video summaries," in *Proc. Int. Conf. Mulmedia and Expo*, 2004, pp. 579-582.

[9] G. Zhu, C. Xu, Q. Huang, Y. Rui, S. Jiang, W. Gao, and H. Yao, "Event Tactic Analysis Based on Broadcast Sport Video," IEEE Trans. Multimedia, vol.11, no.1, 2009, pp.49-67.

[10] S. Chen, M. Chen, C. Zhang, and M. Shyu, "Exciting event detection using multi-level multimodal descriptors and data classification," in Proc. ISM, 2006.

[11] T. Wang, J. Li, Q. Diao, W. Hu, Y. Zhang, and C. Dulong, "Semantic event detection using conditional random fields," in *Proc. Computer Vision and Pattern Recognition Workshop*, 2006, pp.109-115.

[12] N. Nan, G. Liu, X. Qian, and C. Wang, "An SVM-based soccer video shot classification scheme using projection histograms," PCM 2008.
[13] A. Hanjialic, "Generic approach to highlights extraction from a sports video," in *Proc. Int. Conf. Image Processing*, vol.1, pp. 1-4, 2003.
[14] L. Duan, M. Xu, T. Chua, Q. Tian, and C. Xu, "A mid-level representation framework for semantic sports video analysis," in *Proc. ACM Multimedia*, pp. 29-32,2003.
[15] F. Wang, Y. Ma, H. Zhang, and J. Li, "A generic framework for semantic sports video analysis using dynamic Bayesian networks," in *Proc. Int. Conf. Multimedia Modelling*, pp. 29-32,2005.
[16] P. Xu, L. Xie, and S. Chang, "Algorithms and systems for segmentation and structure analysis in soccer video," in *Proc. Int. Conf. Multimedia & Expo*, 2001, pp. 184-187.
[17] C. Xu, J. Wang, H. Lu, and Y. Zhang, "A Novel Framework for Semantic Annotation and Personalized Retrieval of Sports Video," IEEE Transactions on Multimedia, vol. 10, no. 3, 2008, pp.421-436.
[18] L. Duan, M. Xu, Q. Tian, C. Xu, and J. S. Jin, "A unified framework for semantic shot classification in sports video," *IEEE Trans. Multimedia*, vol.7, No.6, 2005, pp.1066-1083.
[19] X. Qian, H. Wang, G.Liu, Z. Li, and Z. Wang, "Soccer Video Event Detection by Fusing Middle Level Visual Semantics of an Event Clip", in Proc. PCM 2010, pp.439-451.
[20] L. Duan, M. Xu, and Q. Tian, "Semantic shot classification in sports video," in *Proc. SPIE Storage and Retrieval for Media Database*, vol. 5021, 2003, pp. 300-313.
[21] X. Qian, Guizhong Liu, Zhe Wang, Zhi Li, Huan Wang, "Highlight Events Detection in Soccer Video using HCRF," in Proc. ICIMCS, 2010.
[22] X. Zhu, X. Wu, A. Elmagarmid, Z. Feng, and L. Wu, "Video data mining semantic indexing and event detection from the association perspective," IEEE Trans. Knowledge and Data Engineering., vol.17, no.5, pp.665-677, 2005.
[23] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. Huang, "Highlights extraction from sports video based on an audio-visual marker detection framework, " in Proc. Int. Conf. Multimedia & Expo, pp. 29-32,2005.
[24] C. Xu, Y. Zhang, G. Zhu, Y. Rui, H. Lu, and Q. Huang, "Using Webcast Text for Semantic Event Detection in Broadcast Sports Video," IEEE Trans. Multimedia, vol.10, no.7, 2008, pp.1342-1325.
[25] Y. Wang, Z. Liu, and J. Huang, "Multimedia content analysis using both audio and video clues," IEEE Signal Processing Magazine, 2000.
[26] J. Assfalg, M. Bertini, C. Colombo, A. Bimbo, and W. Nunziati, "Semantic annotation of soccer videos: automatic highlight identification," Computer Vision and Image Understanding, vol.6, No.4, pp.285-305, Aug.2003.
[27] C. Huang, H. Shih, and C. Chao, "Semantic analysis of soccer video using dynamic Bayesian network," IEEE Trans. Multimedia, vol.8, No.4, pp.749-760, Aug. 2006.
[28] M. Dao, and N. Babaguchi, "Mining temporal information and web-casting text for automatic sports event detection," in Proc. MMSP, 2008, pp.616-621.
[29] M. Dao, and N. Babaguchi, "Sports event detection using temporal patterns mining and web-casting text," in Proc. ACM AREA, 2008, pp.33-40.
[30] X. Qian, and G. Liu, "Global motion estimation from randomly selected motion vector groups and GM/LM based applications," Signal, Image and Video Processing, 2007.

[31] D. Zhang, and S. Chang, "Event detection in baseball video using superimposed caption recognition," in Proc. ACM Multimedia, Juan-les- Pins, France, Nov. 1, 2002, pp. 315-318.

[32] Y. Su, M. Sun, and V. Hsu, "Global motion estimation from coarsely sampled motion vector field and the applications", IEEE Trans. Circuits Syst. Video Technol., Vol. 15, No. 2, Feb. 2005, pp. 232-242.

[33] X. Qian, G. Liu, H. Wang, and R. Su, "Text detection, localization and tracking in compressed videos," Signal Processing: Image Communication, vol.22 , 2007, pp.752-768.

[34] M. Lyu, J. Song, and M. Cai, "A comprehensive method for multilingual video text detection, localization, and extraction," IEEE Trans. Circuits and Systems for Video Technology, vol.15, no.2,2005, pp.243-255.

[35] G. Jin, L. Tao, and G. Xu, "Hidden markov model based events detection in soccer video," ICIAR 2004, LNCS 3221, pp.605-612, 2004.

[36] A. Ekin, A. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," IEEE Trans. Image Processing, vol.12, no.7, 2003, pp. 796-807.

[37] C. Lien, C. Chiang, and C. Lee, "Scene-based event detection for baseball videos," J. Vis. Commun. Image R. 18 (2007) 1–14.

[38] C. Snoek, and M. Worring, "Multimedia event-based video indexing using time intervals," IEEE Trans. Multimedia, vol.7, No.4, pp.638-647, Aug. 2005.

[39] G. Papadopoulos, V. Mezaris, I. Kompatsiaris, and M. Strintzis, "Accumulated motion energy fields estimation and representation for semantic event detection," in Proc. CIVR, 2008, pp.221-230.

[40] F. Wang, Y. Ma, H. Zhang, and J. Li, "Dynamic Bayesian network based event detection for soccer highlight extraction," in Proc. Int. Conf. Image Processing, 2004, pp. 633-636.

[41] D. Sadlier, and N. O'Connor, "Event detection in field sports video using audio-visual features and a support vector Machine," IEEE Trans. Circuits Syst. Video Technol., vol. 15, no. 10, pp. 602-615, Oct. 2005.

[42] K. Wickramaratna, M. Chen, S. Chen, and M. Shyu, "Neural network based framework for goal event detection in soccer videos," in Proc. Int. Symposium on Multimedia. Dec. 2005, pp.21-28.

[43] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257-285, 1989.

[44] C. Cheng, and C. Hsu, "Fusion of audio and motion information on HMM-based highlight extraction for baseball games," IEEE Trans. Multimedia, vol.8, No.3, pp.585-599, June. 2006.

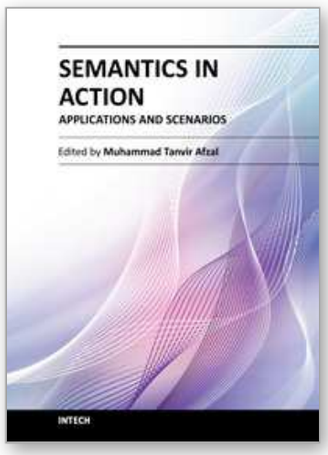
[45] A. Mittal, L. Cheong, and T. Leung, "Dynamic bayesian framework for extracting temporal structure in video," in Proc. Int. Conf. Computer Vision and Pattern Recognition, 2001, pp. 110-115.

[46] N. Dalal, and B Triggs, "Histogram of oriented gradients for human detection," in Proc. Int. Conf. Computer Vision and Pattern Recognition, 2005.

[47] X. Qian, G. Liu, D. Guo, Z. Li, Z. Wang, and H. Wang, "Object categorization using hierarchical wavelet packet texture descriptors," in Proc. ISM 2009, pp.44-51.

[48] C. Sutton, and A. McCallum, "An introduction to conditional random fields for relational learning," Introduction to Statistical Relational Learning. MIT Press. 2007.
[49] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in Proc. Int. Conf. Machine Learning, 2001, pp.282-289.
[50] A. McCallum, "Efficiently inducing features of conditional random fields," in Proc. Uncertainty in Articifical Intelligence, 2003, pp.403-410.
[51] F. Sha, and F. Pereira, "Shallow parsing with conditional random fields," in Proc. Human Language Technology, NAACL, 2003.
[52] M. Mahajan, A. Gunawardana, and A. Acero, "Training algorithms for hidden conditional random fields," ICASSP 2007, vol.1, pp.273-276.
[53] A. Gunawardana, M. Mahajan, A. Acero, and J. Platt, "Hidden conditional random fields for phone classification," in Proc. Int. Conf. Speech Communication and Technology, pp.1117-1120, Sept. 2005.
[54] Y. Sung, C. Boulis, C. Manning, and D. Jurafsky, "Regularization, adaptation, and non-independent features improve hidden conditional random fields for phone classification," in Workshop of Automatic Speech Recognition and Understanding, pp.347-352, 2007.
[55] R. Malouf, "A comparison of algorithms for maximum entropy parameter estimation," in Proc. Int. Conf. National Language Learning, 2002, pp. 49-55.
[56] X. Qian, H. Wang, G. Liu, and X. Hou, "HMM Based Soccer Video Event Detection Using Enhanced Mid-Level Semantic", Multimedia Tools and Applications, 2011. (Accepted)
[57] Z. Liu, and S. Sarkar, "Robust outdoor text detection using text intensity and shape features," in Proc. ICPR 2008
[58] N. Dalal, and B Triggs, "Histogram of oriented gradients for human detection," in Proc. Int. Conf. Computer Vision and Pattern Recognition, 2005.
[59] X. Qian, X. Hua, P. Chen, and L. Ke, "PLBP: An Effective Local Binary Patterns Texture Descriptor with Pyramid Representation", Pattern Recognition, 2011, vol.44, pp. 2502-2515.
[60] X. Wang, and X. Zhang, "ICA mixture hidden conditional random field model for sports event classification," in Proc. ICCV, 2009, pp.562-569.
[61] Y. Tan, D. Saur, S. Kulkarni, and P. Ramadge, "Rapid estimation of camera motion from compressed video with application to video annotation," IEEE Trans. Circuits Syst. Video Technol., vol. 10, no. 1, pp. 133-146, Feb. 2000.
[62] N. Babaguchi, Y. Kawai, and T. Kitashi, "Event based indexing of broadcasted sports video by intermodal collaboration," IEEE Trans. Multimedia, vol.4, no.1, pp.68-75, Mar. 2002.
[63] A. Jain, and B. Yu, "Automatic text location in images and video frames," in Proc. ICPR, 1998, pp. 1497-1499.
[64] V. Mariano, and R. Kasturi, "Locating uniform-colored text in video frames," in Proc. 15th Int. Conf. Pattern Recognit., vol. 4, 2000, pp. 539-542.
[65] X. Qian, and G. Liu, "Text detection, localization and segmentation in compressed videos," in Proc. ICASSP2006., vol. 2, 2006, pp. II385-II388.

[66] T. Sato and T. Kanade, "Video OCR: Indexing digital news libraries by recognition of superimposed caption," ICCV Workshop on Image and Video retrieval. 1998.
[67] Y. Zhong, H. Zhang, and A. Jain, "Automatic Caption Localization in Compressed Video," IEEE Trans. Pattern Analysis and Machine Intelligence, vol.22, no.4, 2000, pp.385-392.
[68] C. Ngo, C. Chan, "Video text detection and segmentation for optical character recognition,". Multimedia Systems, vol.10, no.3, 2005, pp.261-272.
[69] J. Zhang, D. Goldgof, and R. Kasturi, "A New Edge-Based Text Verification Approach for Video," in Proc. ICPR, 2008.
[70] L. Sun, G. Liu, X. Qian, D. Guo, "A Novel Text Detection and Localization Method Based on Corner Response," in Proc. ICME 2009.
[71] U.Gargi, S.Antani, and R. Kasturi, "Indexing text events in digital video databases," in Proc. Int. Conf. Pattern Recognit., vol. 1, 1998, pp.916-918.
[72] X. Tang, B. Gao, J. Liu, and H. Zhang, "A spatial-temporal approach for video caption detection and recognition," IEEE Trans. Neural Networks.,vol. 13, no.4, 2002, pp. 961-971.
[73] H. Jiang, G. Liu, X. Qian, et al., "A Fast and Efficient Text Tracking in Compressed Video," in Proc. ISM 2008.



Semantics in Action - Applications and Scenarios

Edited by Dr. Muhammad Tanvir Afzal

ISBN 978-953-51-0536-7
Hard cover, 266 pages
Publisher InTech
Published online 25, April, 2012
Published in print edition April, 2012

The current book is a combination of number of great ideas, applications, case studies, and practical systems in the domain of Semantics. The book has been divided into two volumes. The current one is the second volume which highlights the state-of-the-art application areas in the domain of Semantics. This volume has been divided into four sections and ten chapters. The sections include: 1) Software Engineering, 2) Applications: Semantic Cache, E-Health, Sport Video Browsing, and Power Grids, 3) Visualization, and 4) Natural Language Disambiguation. Authors across the World have contributed to debate on state-of-the-art systems, theories, models, applications areas, case studies in the domain of Semantics. Furthermore, authors have proposed new approaches to solve real life problems ranging from e-Health to power grids, video browsing to program semantics, semantic cache systems to natural language disambiguation, and public debate to software engineering.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Xueming Qian (2012). Semantic Based Sport Video Browsing, Semantics in Action - Applications and Scenarios, Dr. Muhammad Tanvir Afzal (Ed.), ISBN: 978-953-51-0536-7, InTech, Available from: http://www.intechopen.com/books/semantics-in-action-applications-and-scenarios/semantic-based-sport-video-browsing

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen