Semantic Units Based Event Detection in Soccer Videos[1)]

TONG Xiao-Feng LIU Qing-Shan LU Han-Qing JIN Hong-Liang

(*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing* 100080)

(E-mail: {xftong, qsliu, luhq, hljin}@nlpr.ia.ac.cn)

Abstract A semantic unit based event detection scheme in soccer videos is proposed in this paper. The scheme can be characterized as a three-layer framework. At the lowest layer, low-level features including color, texture, edge, shape, and motion are extracted. High-level semantic events are defined at the highest layer. In order to connect low-level features and high-level semantics, we design and define some semantic units at the intermediate layer. A semantic unit is composed of a sequence of consecutives frames with the same cue that is deduced from low-level features. Based on semantic units, a Bayesian network is used to reason the probabilities of events. The experiments for shoot and card event detection in soccer videos show that the proposed method has an encouraging performance.

Key words Event detection, semantic unit, video semantic analysis, Bayesian network

1 Introduction

With the increasing of multimedia data, it is crucial to find an efficient way to manage the media data, including browse, filtering and retrieval[1~3]. Low-level features are too oversimple to use for semantic requirement. Recently, event based multimedia indexing and retrieval is widely concerned[4~6], and it is much more significant and valuable than shot based video analysis. Generally speaking, an event can be regarded as an interesting activity in a video segment, and it should have the three basic characteristics: 1) domain-dependent; 2) spatial-temporal context related; 3) difficult to be simply characterized and identified by low-level features. This paper focuses on semantic event detection and analysis in soccer programs. For a lengthy soccer game, highlights often take up a small part, so it is very significant to detect and analysis events in soccer video.

At present, some studies have been done on the event detection and analysis in sports video. Naphade *et al.*[7] presented concepts of multi-objects and multi-nets, and set up a multi-nets framework based on graph probabilistic reasoning for semantic video indexing. A general "event + non-event" framework for indexing and summarizing sports broadcast programs was presented in [8]. Vasconcelous *et al.*[9] put forward a Bayesian framework to extract video semantic features to depict content of movies, but their method did not consider temporal context. In [10], a scene detection and structure analysis method for sports video was developed, which combined domain-specific knowledge, supervised machine learning and hierarchical features analysis technology. P. Xu *et al.*[11] developed a method that divided a sports video into play and break segments. Based on this work, L. Xie *et al.*[12] employed HMM and dynamic programming to enhance the performance of segment detection and classification with taking field-ratio and motion activity as observations. [13] analyzed video editing ways and object based features. They proposed an automatic soccer program analysis and summarization method. In their experiments, they detected slow-motion replay, close-up, break, and utilized a heuristic rule to identify highlights. X. Sun *et al.*[14] used Bayesian network to detect score events based on goalnet, audience, scoreboard and face cues.

In this paper, we propose a semantic unit based event detection scheme according to the characteristics of events in sports video. The scheme can be characterized as a three-layer framework shown in Fig. 1. At the lowest layer, low-level features, such as color, texture, edge, shape and motion, are extracted from visual frames. Events describing interesting activities in video segments are defined at the highest layer. In order to bridge low-level features and high-level semantic events, we define semantic units at the intermediate layer. A semantic unit is composed of continuous frames that contain the same cue. Semantic units are derived from low-level features and taken as observation of event inference. Generally, an event consists of several semantic units. Presence of some specific semantic units

1) Supported by National Natural Science Foundation of P. R. China (60475010, 60121302)
Received January 14, 2004; in revised form September 17, 2004

indicates a specific event. In our experiments, we employ this scheme to detect shoot and card events in soccer videos. Considering the domain knowledge, we define six types of semantic units: replay, goalmouth, caption, close-up, audience, and close+caption unit. Taking these units as observations, a Bayesian network is to reason the probabilities of defined events.

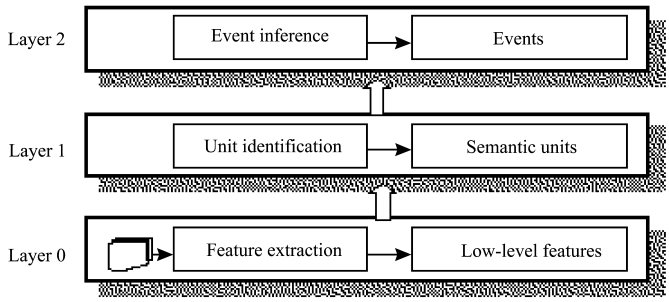

Fig. 1 Framework of the scheme

The rest of the paper is organized as below. Section 2 introduces low-level features. Section 3 discusses detection of semantic units. Section 4 describes event inference. Experiments are given in Section 5. Conclusions are drawn in Section 6.

2 Low-level features

Low-level features include field dominant color, skin color, frame-to-frame difference, edge, texture, shape of region, and scale of objects in the field.

1) Field dominant color: Game field extraction is an important procedure in event detection. To reduce the effect of illumination, we select HSV color space, and only use hue and saturation components. Assuming H_{mean} and S_{mean} the values of hue and saturation components of filed dominant color, they can be obtained through statistics at the start period of the game[13]. The distance from a pixel $f(i,j)$ to the dominant color values is defined as below.

$$d_{hsv} = \sqrt{S^2(i,j) + S^2_{mean} - 2S(i,j)S_{mean}\cos(\theta)}$$

where $\theta = |H(i,j) - H_{mean}|$, $H(i,j)$ and $S(i,j)$ are hue and saturation components of the pixel $f(i,j)$. If the distance is smaller than a threshold, this pixel belongs to the field.

2) Skin detection: An effective unimodal Gaussian model with multi-variable is utilized to detect skin region[15]. Then, morphological filtering is applied to remove small and crash areas. The shape and scale of the skin area are used to identify close-up views.

3) Frame-to-frame difference: The mean square difference (MSD) of intensity is used to measure the difference of adjacent frames. MSD is utilized to detect logo-transitions in replay segments.

4) Edge: Edge is also a very useful feature. We apply a Sobel operator with the size of 3×3 to extract edges of an image. Edge information is used to discriminate goalmouth and caption area.

5) Texture: Texture describes the repeated mode of local changes of image intensity, and it often takes the gray spatial distribution of neighbors of pixels as features. It is utilized to distinguish audience from out-field close-up views.

6) Shape: Shape is used for verifying head area after shin detection. Shape feature includes: 1) scale, *i.e.*, the height of region; 2) compactness, *i.e.*, ratio of actual area to the area of the min-bounding-box; 3) elongation, *i.e.*, ratio of height to width of the min-bounding-box.

7) Scale of objects in field: It is defined as the ratio of average height of objects to that of game field in the frame. It directly reflects the distance from camera to the captured objects. Before scale estimation, object in field segmentation is necessary. For detailed algorithm, please refer to our previous work[15].

3 Semantic units

A semantic unit is composed of continuous frames with the same cue. It is a descriptor for a video segment, and bridges the low-level features and high-level semantics. In sports video, selection

of semantic units usually needs to consider domain-dependant knowledge and video editing rules. We concern shoot and card events in soccer videos in this paper. Correspondingly, we define six types of units: replay, goalmouth, caption, close-up, audience, and close+caption units. An interesting shoot event usually contains replay, goalmouth and player close-up units. Furthermore, a scene of excited audience and scoreboard will appear if score. A serial of typical views in a score event are shown in Fig. 2. In a server foul event, such as red/yellow card event, a red/yellow card recorder will be superimposed onto the player close-up views in addition to replay segment.

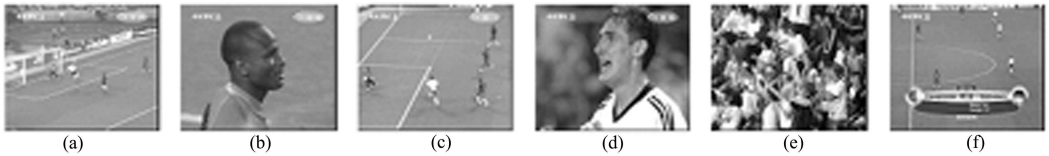

Fig. 2 Typical views of a goal event
(a) Goalmouth, (b) Close-up, (c) Replay, (d) Close-up, (e) Audience, (f) Scoreboard

The operation of semantic units is carried out on frames. If the counter of continuous frames containing the same cue exceeds a threshold, a semantic unit is declared. Semantic unit detection is kept in the following order:

Step 1. Replay segments detection.

Further processing in the rest segments apart from replays in the following order:

Step 2. Caption detection.

Step 3. Views classification, obtain close-up and audience view.

Step 4. Based on step 2 and 3, identify close+caption views.

Step 5. Detect goalmouth in long views.

1) Replay. Replay is a video editing way, and it is often used to emphasize an important segment with a slow-motion pattern for once and several times. At present, there are two methods for replay detection, *i.e.*, adjacent frame difference based method[16] and compressed prediction vectors based method[17]. They are valid for some replay segments generated by special means. In this paper, we apply a simple and effective detection method based on replay-logo.

In sports video, there is often a highlighted logo that wipes at the start and end of a replay segment and the logo is invariant in the whole video. Therefore, we can firstly obtain the logo from these wipe transitions and then employ it to detect replay segments. The replay detection algorithm[15,18] consists of the following steps: 1) Detect no less than n logo-transitions and extract an optimal candidate of logo in each of them. 2) Take these candidates as a cluster and get its center. Compute the mean image of those candidates near to the center to eliminate the effect of background. The mean image is then regarded as the logo template. 3) Extract other logos through the logo template matching in the video. A pair of logos determines a replay segment. A logo-transition and extracted logo are displayed in Fig. 3.

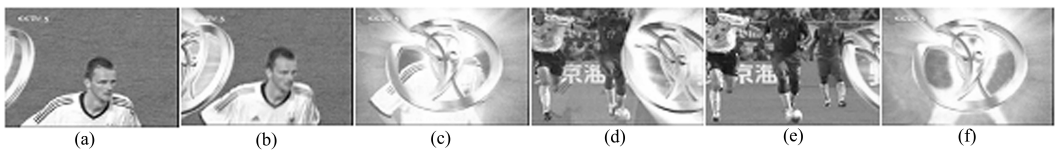

Fig. 3 Five images in a transition (a∼e) and a logo-template image (f)

2) **Caption.** In soccer videos, caption is appeared at these cases: recorder score, red/yellow card, player substitution and technical statistics. It is difficult to recognize the text in a caption, such as player names, score, but the appearance of caption usually indicates an occurrence of special event.

The caption region can be treated as a special texture area aligned by vertical strokes, in which the gradients of local neighbors are greater and more uniform than those of other regions. The procedure of caption area detection[19] consists of gradient computation, run-length smoothing, morphological open

operation, region segmentation and verification. Because captions are often appeared at the bottom of an image, we just need to do such detection at the bottom of frames.

3) Close-up and audience. The focal players are attracted with close-up view in a highlight segment, such as shoot and card events. In red/yellow card events, close-up views usually are superimposed upon the caption of card recorder. In shoot events, views of excited audience will also be shown. We utilize a decision tree to classify views into long, medium, close-up or audience type based on field-ratio, texture, qualified head area and object scale in game field[15].

4) Close+Caption. When caption appears in a close-up frame, we treat it as a close+caption view independently. Close+caption views usually appear in the case of server foul or players substitution.

5) Goalmouth. Goalmouth is also a valid cue for highlights. Fig. 4 (a) shows a long side view in a shoot segment. A goalmouth is composed of a goal line, goal posts and a crossbar. We restrict the region of edge detection to reduce noise. The detection procedure includes: 1) Compute the coarse spatial representation CSR (i, j) of the original image, shown in Fig. 4 (b)[20]. 2) Extract edges in the region between field and non-field in CSR, shown in Fig. 4(c). 3) Search the longest line in the edge image, $L(\rho, \theta)$. In common, the slant angle of the goal line in the image captured by the main camera (placed at near the middle of game field) is relatively fixed. So, we can define an interval to restrict the angle of the goal line and filter false alarms. 4) Goal posts and crossbar detection based on gray growing after the goal line extraction. The final result is shown in Fig. 4 (d)

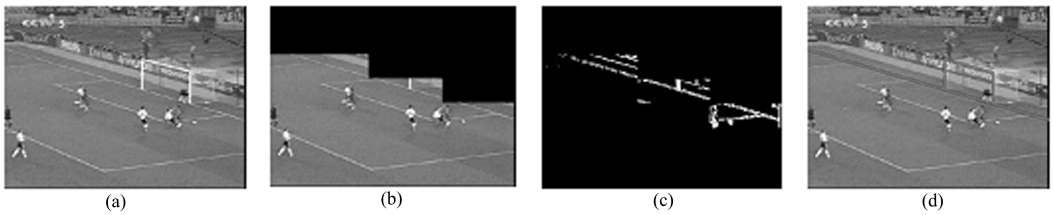

Fig. 4 Goalmouth detection. (a) Original image, (b) CSR, (c) Edge in CSR, (d) Result (overlay with red line)

6) Video decomposition based on semantic units. According the above definition and discussion, we can partition a video into a sequence of semantic units. Combination of special semantic units indicates the presence of a special event. Fig. 5 gives the comparison of semantic units and shots based video decomposition in a video. The upper seven rows are timelines of semantic units, and every horizontal red line segment denotes a semantic unit. The bottom rows describe the video decomposition based on shots.

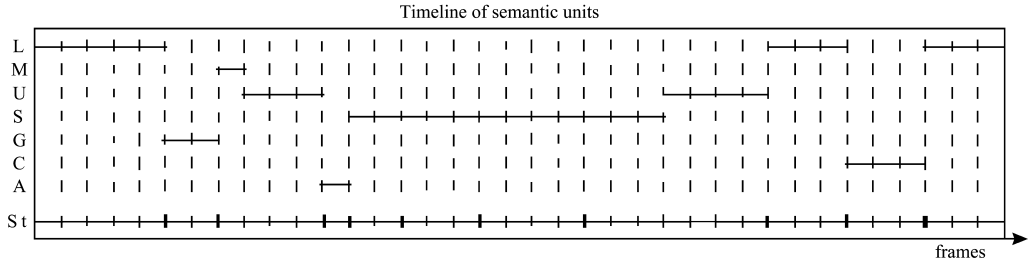

Fig. 5 Semantic units representation of a video clip. L – long view unit, M – medium view unit, U – close-up view unit, S – SMR unit, G – goalmouth unit, C – caption unit, A – audience unit; St - shot

4 Event inference

At the highest layer of the framework, a Bayesian network is used to reason the probabilities of presence of events with semantic units as observations. Bayesian networks are directed acyclic graphs (DAGs) representing the causal dependencies between the nodes that hold variable[21]. Bayesian network reasoning is the procedure of computing the posterior probabilities of required objective nodes

using prior probabilities in conditional probabilities dataset and known nodes. The correlation between observations and conclusions can be measured by mutual information.

In this paper, we construct a Bayesian network shown in Fig. 6 to detect shoot and card events in soccer videos. For shoot event, replay, audience, goalmouth, caption and close-up units are taken as observations. For red/yellow card event, close+caption unit replaces caption and close-up unit.

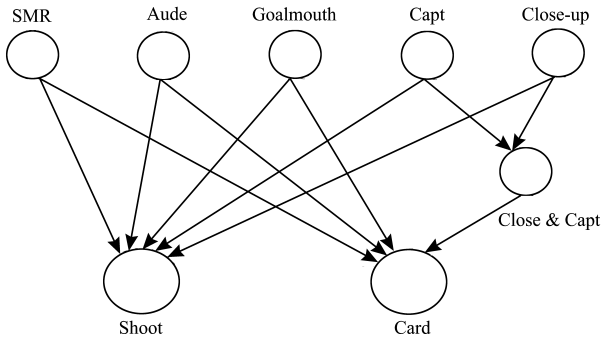

Fig. 6 Structure of the Bayesian network

5 Experiments

We apply the proposed scheme to detect shoot and red/yellow card events in real soccer videos. The model parameters of prior and conditional probabilities are obtained by statistics over a dataset of 200 clips, totally about 450 minutes. The test dataset has 32 clips. All of these data are manual elaborately grabbed from the FIFA World Cup 2002. In the testing clips, 17 of them are shoot events, 12 are card events, and the rest clips are uninteresting events. The ground truth events are labeled manually.

Table 1 reports the performance of event detection of our method. For shoot events, 14 of them are correctly detected, 3 of them are missing, no false detection. In the 3 missing clips, 2 clips have no player close-up views, and one clip does not appear replay unit. For card events, 10 of 12 are successfully detected, 2 of them are missing, no false results. In the 2 missing clips, 1 clip has no close+caption unit, and the other one has no replay unit. Perhaps because the training data is not enough, we cannot learn the accurate network model for these two events detection.

Table 1 Results of event detection

Event	Truth	Detect	False	Miss
Shoot	17	14	0	3
Card	12	10	0	2

The results of semantic units detection is shown in Table 2. In our experiments, the detection operation is carried out on every three frames. The final classification of unit is determined by votes of frames in a segment. When the count of frames containing the same cue exceeds a certain threshold, a semantic unit is declared. In unit detection, the performance of close-up unit detection is not good. Most of false alarms are due to wrong classification for medium views. In total, the performance of units′ generation is reliable, which guarantees satisfied detection of events.

Table 2 Results of semantic units detection

Se-Unit	Truth	Detect	False	Miss
SMR	26	26	0	0
Aude	4	4	0	0
Goal	27	26	1	1
Close	53	47	10	6
Capt	16	16	0	0
Co+Cp*	13	13	0	1

*Se_Ut—Semantic Unit; Ca+Cp—Close & Capt

6 Conclusions

In this paper, a semantic unit based event detection scheme is proposed. It can be characterized as a three-layer framework. At the lowest layer, low-level features including color, texture, edge, shape and motion are extracted. Semantic events are defined at the highest layer. In order to bridge the semantic gap between low-level features and high-level events, we define some semantic units as domain-dependent knowledge and specific interesting events at the intermediate layer. A semantic unit is composed of consecutive frames containing the same cue. It is a descriptor for a video segment. A Bayesian network based probabilistic framework is used to event inference with the semantic units as observations. We apply this scheme to detect shoot and card events in soccer videos. The experiments demonstrate the validity and effectiveness of this scheme.

In the future, the dynamic probabilistic networks (DPNs) analysis for time sequence and automatic video segmentation based on events are should be studied.

References

1 Zhang H, Wang A, Altunbasak Y. Content-based video retrieval and compression: A uinified solution. In: Proceedings of IEEE International Conference on Image Processing. Santa Barbara, USA: IEEE Computer Society Press, 1997. **1**: 13∼16

2 Zhong D, Chang S-F. Spatio-temporal video search using the object-based video representation. In: Proceedings of IEEE International Conference on Image Processing. Santa Barbara, CA, USA: IEEE Computer Society Press, 1997. **1**: 21∼24

3 Arman F, Depommier R, Hsu A, Chiu M Y. Content-based browsing of video sequences, ACM International Conference on Multimedia, San Francisco, USA: ACM Press, 1994. 97∼103

4 Haering N, Qian R, Sezan M. A semantic event-detection approach and its application to detecting hunts in wildlife video. *IEEE Transactions on Circuits and Systems for Video Technology*, 2000, **10**(6): 857∼868

5 Yow D, Yeo B, Yeung M, Liu B. Analysis and presentation of soccer highlights from digital videos. In: Proceedings of the 2nd Asian Conference on Computer Vision. Singapore: Spring Press, 1995. 499∼503

6 Gong Y, Sin L-T, Chuan C-H, Zhang H-J, Sakauchi M. Automatic parsing of TV soccer programs. In: Proceedings of IEEE International Conference on Multimedia Systems and Computing, Washington DC, USA: IEEE Computer Society Press, 1995. 167∼174

7 Naphade M-R, Huang T-S. Semantic video indexing using a probabilistic framework. In: Proceedings of IAPR International Conferecne on Pattern Recognition. Barcelona, Spain: IEEE Computer Society Press, 2000. **3**: 83∼88

8 Li B, Errico J. Bridging the semantic gap in sports. In: Proceedings of IS & T/SPIE Conference Storage and Retrieval for Median Databases. Santa Clara, California, USA: SPIE Press, 2003. **5021**: 314∼326

9 Vasconcelos N, Lippman A. A Bayesian framework for semantic content characterization. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. Santa-Barbara: IEEE Computer Society Press, 1998. 566∼571

10 Zhong D, Chang S-F. Structure analysis of sports video using domain models. In: Proceedings of IEEE International Conference on Multimedia and Expo, Tokyo, Japan: IEEE Computer Society Press, 2001. 713∼716

11 Xu P, Xie L, Chang S-F, Divakaran A, Vetro A, Sun H. Algorithms and systems for segmentation and structure analysis in soccer video. In: Proceedings of IEEE International Conference on Multimedia and Expo, Tokyo, Japan: IEEE Computer Society Press, 2001. 721∼724

12 Xie L, Chang S-F, Divakaran A, Sun H. Structure analysis of soccer video with hidden Markov models. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Orlando, FL, USA: IEEE Computer Society Press, 2002. **4**: 4096∼4099

13 Ekin A, Tekalp A M, Mehrotra R. Automatic soccer video analysis and summarization, *IEEE Transactions on Image Processing*, 2003, **12**(7): 796∼807

14 Sun X, Jin G, Huang M, Xu G. Bayesian network based soccer video event detection and retrieval. In: Proceedings of SPIE International Conference on Multi-spectrum Image Processing and Pattern Recognition. Beijing, China: SPIE Press, 2003. **5286**: 71∼76

15 Tong X F, Liu Q S, Lu H Q, Jin H L. Shot classification in sports video. In: Proceedings of the 7th International Conference on Signal Processing, Beijing, China: IEEE Computer Society Press, 2004, 1364∼1367

16 Pan H, Beek P, Sezen M. Detection of slow-motion replay segments in sports video for highlights generation. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, USA: IEEE Computer Society Press, 2001, 6149∼1652

17 Kobla V, Dementhon D, Doermann D. Detection of slow-motion replay sequences for identifying sports videos. In: Proceedings of IEEE Signal Processing Society 1999 Workshop on Multimedia Signal Processing, Copenhagen, Denmark: IEEE Computer Society Press, 1999. 135∼140

18 Tong X F, Lu H Q, Liu Q S. A three-layer framework of event detection in soccer videos. In: Proceedings of IEEE International Conference on Multimedia and Expo, Taiwan: IEEE Computer Society Press, 2004. 1551∼1554

19 Wolf C, Jolin J, Chassaing F. Text localization, enhancement and binarization in multimedia document. In: Proceedings of IAPR International Conferecne on Pattern Recognition, Quebec, Canada: IEEE Computer Society Press, 2002. **2**: 1037∼1040

20 Wan K, Yan X, Yu X, Xu C. Real-time goal-mouth detection in MPEG soccer video. In: Proceedings of ACM International Conference on Multimedia, USA: ACM Press, 2003. 311∼314

21 Charniak E. Bayesian networks without tears. *AI Magazine*, Winter 1991, 50∼63

TONG Xiao-Feng Received his bachelor degree from School of Information Engineering, China University of Geosciences in 1999; master degree from Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology in 2002. He is current a Ph. D. candidate in National Laboratory of Pattern Recognition at Institute of Automation, Chinese Academy of Sciences. His research interests include video analysis, multimedia indexing, and pattern recognition.

LIU Qing-Shan Received his master degree from Southeast University in 2000, the Ph. D. degree from National Laboratory of Pattern Recognition. He is now an assistant professor of Institute of Automation, Chinese Academy of Sciences. His research interests include pattern recognition, machine learning, and image and video analysis.

LU Han-Qing Received his bachelor and master degrees form Harbin Institute of Technology in 1982 and 1985, respectively, the Ph. D. degree from Huazhong University of Science and Technology in 1992. He is now a professor of Institute of Automation, Chinese Academy of Sciences. His research interests include image and video processing, multimedia indexing and retrieval.

JIN Hong-Liang Received his bachelor and master degrees from North China Electric Power University in 1999 and 2002, respectively. He is now a Ph. D candidate in National Laboratory of Pattern Recognition at Institute of Automation, Chinese Academy of Sciences. His research interests include face detection, face recognition, and machine learning and pattern recognition.