SEMANTIC UNITS BASED EVENTS DETECTION IN SOCCER VIDEOS

Xiaofeng Tong, Qingshan Liu, Hanqing Lu
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
P.O.Box 2728, Beijing, 100080, China
{xftong, qsliu, luhq}nlpr.ia.ac.cn

ABSTRACT

In this paper, an event detection scheme based on semantic units is proposed. The scheme can be characterized as a three-layer framework. At the lowest layer, low-level features including color, texture, edge, shape, motion, etc. are considered. High-level events are modeled and characterized at the highest layer of this framework. In order to bridge the semantic gap between low-level features and high-level semantics, we define some semantic units at the intermediate layer. A semantic unit, which is composed of consecutive frames with similar cue, is a description for video segment. It is deduced by frame-wise low-level features and taken as evidence of event reasoning. In event inference, a Bayesian network based on probabilistic framework is utilized. Experiments have demonstrated that the scheme is effective and valid.

1. INTRODUCTION

With the increasing amount of multimedia, there is an emerging need for efficient media management including browsing, filtering, indexing, and retrieval. One of the big challenges facing this task is the problem of bridging the semantic gap between low-level features and video semantics. A lot of efforts have been done to solve the problem. Many of them are carried out on sports video for its relative regular structure and enormous potential applications.

In recent years, Naphade et al [1] applied a probabilistic framework to build a multinet to enhance the detection performance of multijects and integrate multiple modalities to infer unobservable concepts. A general event+non-event framework for indexing and summarizing sports broadcast programs was presented in [2]. But their definition of event just refered to highlights for continuous action sports. Xie et al [3] decomposed a soccer video into two states: play and break based on grass-ratio and motion intensity with hidden Markov models. Ekin et al [4] proposed a system for automatic soccer video analysis and summarization using cinematic and object-based features. But their slow-motion replay detection algorithm was not robust at some cases. [5] performed regions classification but did not have any component to actually model the temporal dynamics. [6] proposed a Bayesian net to model the structure but the network was not trained from samples and conditional densities were fixed heuristically.

Two key issues should be considered in event detection: 1) event modeling; and 2) features (low- and intermediate- level) selection and extraction. In the first issue, a reasonable model should be put forward to describe and characterize an event. It can approximately explain the logical and casual relations between the event and its observations. In the latter one, valid and effective low-level and intermediate-level features are selected and extracted to be evidences for event reasoning. Many efforts have been done on event detection.

In this paper, we propose a three-layer scheme for event detection. This scheme embodies hierarchical characteristics of video content. Different layer of the framework describes different level's video object. The lowest layer depicts a frame with low-level features. The intermediate layer characterizes a video segment with semantic cue. And the highest layer usually represents high-level semantic event that takes place in a video clip. The hierarchical model facilitates video analysis and event detection. At the lowest layer, we select and extract some useful and effective visual features. It is a task-driven procedure, which is essential in pattern recognition. An event at the highest layer describes a video clip with high-level semantics. It is usually related to the domain to be processed. Generally, there is a gap between low-level features and high-level semantics. It is just the classic problem in information retrieval. In order to bridge the gap, we define some semantic units at the intermediate layer according to the properties of high-level events. In the definition, we often take into account domain-depend knowledge and video editing way to facilitate event detection. A semantic unit describes a video segment with certain semantics. It is deduced by low-level features and servers as evidence of high-level events. In event inference, a Bayesian network is utilized to probabilistic reasoning. In design and selection of semantic units, we taken into account the inherent video information and posterior video editing rules, both of them are useful cues for event reasoning.

To test out scheme, two types of events are considered in this paper: shoot and red/yellow card. Consequently, six classes of semantic units are selected: SMR unit, goalmouth unit, caption unit, close-up unit, audience unit, and close-up with superimposition upon caption (Close&Caption) unit. All of them are important observations for the two types of events.

In Section 2 we describe low-level features. Semantic units generation is developed in Section 3. Event inference is presented in Section 4. Experiments and discuss are given in Section 5. Finally, Conclusions are drawn in Section 6.

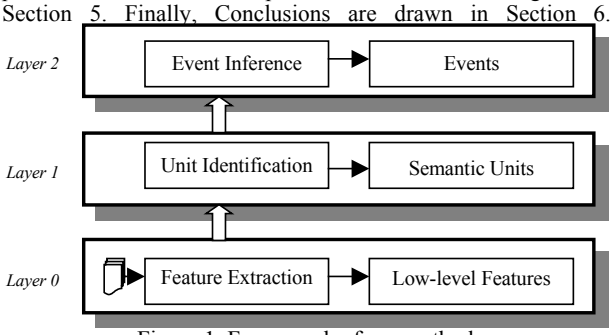


Figure 1. Framework of our method

2. LOW-LEVEL FEATURES

The low-level features include some common and custom-built visual features, which serve as input of semantic unit identification.

For specific event detection and analysis, the required features are carefully selected according to definition and characteristics of event. The low-level features in this paper include color, edge, texture, shape, etc. Though motion is important for video

analysis, it is not effective in SMR analysis at present. Therefore, we extract SMR unit via replay-logo transition identification.

2.1. Color

For color information, we concern field dominant color and skin color in this paper. The field dominant color is used to extract game field in a soccer view. Skin color is utilized in face detection in view classification.

A soccer field has one distinct dominant color, which can be obtained through statistics over a large quantity of field colored images. In computation, only hue and saturation components in HSV color space are used. Set the value of hue and saturation of field be H_{mean} and S_{mean}, respectively. The distance from a pixel $f(i,j)$ to the field dominant color is measured by cylinder metric.

$$\theta = \left| H(i,j) - H_{mean} \right|$$

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

$H(i,j)$ and $S(i,j)$ are hue and saturation components respectively. The field region is composed by these pixels whose $d_{hsv} < TH$, TH is a threshold. Figure 2 is an example. We can see that the segmentation result is perfect even if there are shadows.

Figure 2. Original image (left), and field extraction (right).

Skin area is used in head area detection in the module of close-up view identification.

A simple and effective unimodal Gaussian model with multi-variables is utilized to detect skin area [7].

2.2. Edge

To capture edges, a Sobel operator with a 3*3 window is applied to the original image and the edge map is obtained. Edge detection is employed in gradient computation in caption detection.

2.3. Texture

Gray-level co-occurrence matrix (GLCM) of the intensity is computed and a statistical feature of GLCM, contrast, is calculated. The contrast of GLCM is a measure of spread of the matrix values and smoothness of pixels variation in their local neighborhood. Texture is used in view type discrimination.

2.4. Shape

Shape features are used in verification of head area after skin detection. The involved features include: 1) scale, the height of detected area; 2) solidity, ratio of its area to the min-max box; and 3) orientation, ratio of height to width of the min-max box.

2.5. Scale of Object

Scale of object is employed in view classification. It is defined as the ratio of object's height to field's height in a view. Therefore, we should firstly segment objects (players mostly) in field and estimate their height. Please refer to [9] for detailed shape and scale features' extraction.

3. SEMANTIC UNITS

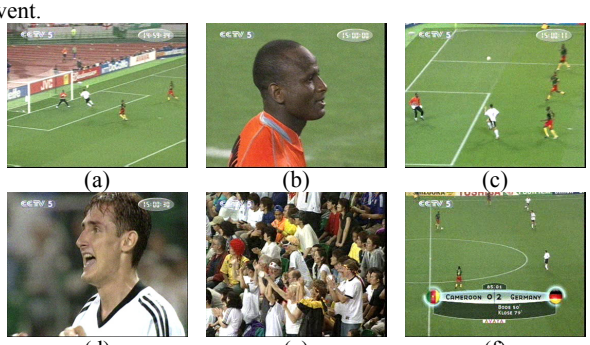
A semantic unit is a sequence of consecutive frames embodying certain semantics. It is a intermediate description for a video segment. In sports video, the definition and property of semantic unit is usual relative to domain and video editing rules.

In this paper, we concern shoot and card events in soccer video to test our scheme. We define six types of semantic units: slow-motion replay (SMR), goalmouth, caption, close-up, audience and close-up with superimposition upon caption (Close&Caption) units. The unit detection operation is carried out on frames. If the counter of consecutive frames belonging to the same type of unit exceeds a certain threshold, the corresponding semantic unit is declared.

The semantic units are detected as the following order:

Step 1 SMR unit recognition with replay-logo.
In the rest frames except those in SMR unit, we do
Step 2 Caption detection in all frame and labeling.
Step 3 View classification [9], we then get long, medium, close-up and audience views.
Step 4 Close&Caption view identification based on step 2 and 3.
Step 5 Goal post slant angle estimation in long view to identify goalmouth view.

A complete score event in soccer game usually contains goalmouth, close-up, SMR, audience cheer and caption (display score) units. These views are displayed in Figure 3. In practice, it is unnecessary that all above view should present in a shoot event.



(a) (b) (c)

(d) (e) (f)

Figure 3. A complete score event. (a) a goalmouth view; (b)a close-up for goalkeeper; (c) SMR; (d) a close-up for shooter; (e) excited audience, and (f) caption for score

3.1. SMR Unit

SMR is an important video editing way that is often used to replay the important segments at a slow speed. It is often taken as a significant cue for event detection.

It is known to us that usually there are logo transitions both at the start and end of replay segment. A pair of logo transitions determines a replay segment. Till now, there are no robust and effective SMR detection methods. To avoid complex direct detection of SMR [7], we present an effective and valid SMR detection algorithm based on replay-logo.

The algorithm consists of the following steps: 1) Automatic detection of logo-transition. If the counter of consecutive frame-to-frame difference of intensity that exceeds a threshold is grater than certain number, a qualified candidate of logo- transition is declared. 2) Extraction of logo template. Take the optimal frame under some conditions as a candidate of logo from logo-transition. The above two steps are carried out on the whole video in advance. 3) Logo detection in video clip by logo template. The similarity is measured by both color and shape. A pair of logo transition determines a SMR. For detailed procedure, please refer to our previous work [9].

A SMR unit is individual segment; no other detection operations are performed on the frames within it.

3.2. Goalmouth Unit

Goalmouth is also a valid cue for highlights. It is usually displayed in a long shot before shoot. Generally, in a goalmouth view, field-ratio is high and the slant angle of goal post exist a certain range (Figure 4). The above two properties almost can characterize a goalmouth view which is usually captured by cameras placed in a near-fixed position. Thus, we detect a goalmouth via computation of field-ratio and estimation of slant angle of goal post.

Firstly, we segment game field and compute the field-ratio. Field extraction is discussed in section 2.1.

If the field-ratio exceeds certain threshold, we further estimate the slant angle of goal post through line fitting (Hough transform is equivalent at this case). If the slant angle is within a certain range, a goalmouth view can be declared. The result is displayed in Figure 4.

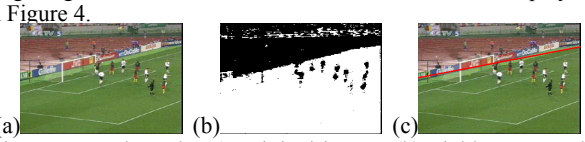


Figure 4. Goalmouth. (a) Original image; (b) Field segmented image; (c) Slant angle estimation (overlay with a red line)

3.3. Caption Unit

Caption indicates content or compensatory information of scene in videos. There are two types of caption: 1) Manual-label caption, it is superimposed upon the original video stream via posterior video edit. Manual-label caption contains the semantic description for current video content. 2) Scene caption, it is naturally embedded in objects or environment in a view, such as words on clothes, trademark of product, etc. Though it indicates some information, it is difficult to find out their common characteristics to analysis and recognition. So in this section, only manual-label caption is concerned.

Text is treated as a special texture aligned by vertical strokes; in text region the gradients of local neighbors are greater and more uniform than that of other regions. Caption region is detected by local-accumulated gradient [10], which consists of gradient computation (ref. Section 2.2), run-length smoothing, morphological open operation, region segmentation and region verification. Only the bottom of the screen needs to be examined in sports video.

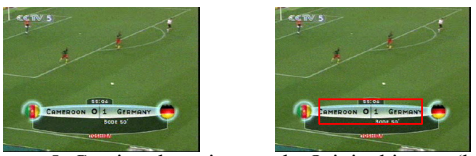


Figure 5. Caption detection result. Original image (left) and detection result (marked by a red rectangle)

3.4. Close-up and Audience Unit

A close-up view is often used to give prominence to a player whose action is fascinating. A cheer scene of audience often display after a highlight.

We classify a view into long, medium, player close-up or audience view (SMR is excluded here) by using a decision tree with some discriminating features, such as field-ratio, texture, head area and object-scale. Note that there are two sub-types of close-up view, close-up with field background (Figure 6(a)) and close-up with non-field background (Figure 6(b)). Both of them belong to the same type, but have different view appearance. The detailed algorithm was discussed in [9].

Figure 6. Close-up view (a) and (b); audience view (c)

3.5. Close&Caption Unit

If a close-up view contains caption simultaneously, it is declared as a close&caption view. These views usually appear when serious foul, such as red or yellow card events. For recognition of Close&Caption view, we firstly identify whether a view contains caption or not. If yes, we further detect close-up features on it.

4. EVENT INFERENCE

At the highest layer, taking the semantic units as observations, a Bayesian network is used to reason the probabilities of interested events. Bayesian networks are directed acyclic graphs (DAGs) representing the causal dependencies between the nodes that hold variables [11]. They have an ability to respond to changing conditions easily and a strong property in causal reasoning that is necessary to model action, explanations, counterfactuals and preferences. The knowledge of the domain is used to construct the network. The inherent uncertainty in the evidences that can be controlled from spatio-temporal data set is represented by the prior probabilities of some variable and conditional probabilities between the variables. When the evidence is observed, it is inserted to the network and the posterior probabilities of event are calculated using model parameters, priors and conditional probabilities.

In the procedure of construction of a Bayesian network, prior knowledge and domain-dependent rules are considered. Some useful and effective observations for reasoning are selected. The structure of Bayesian network in this method is shown in Figure 7. For shoot event, five observations: SMR, audience, goalmouth, caption and close-up unit, are extraordinary concerned. For yellow/red card event, close&caption unit replaces the caption and close-up units. In practice, the experience can guide us to design the structure of Bayesian network. Maybe we can give the probable rank of the reasoning probabilities on edges from observations to results according to our experience, but in experiments, these probabilities should be obtained through large training data set.

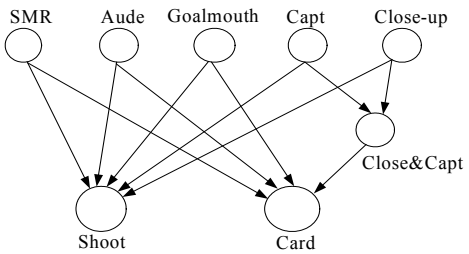


Figure 7. Bayesian network for event detection

5. EXPERIMENTS AND DISCUSS

We detect shoot and yellow/red card events in real soccer video with this scheme. The model parameters of prior and conditional probabilities are obtained by training over a data set of 200 clips, totally about 450 minutes. The clips are segmented and labeled manually. Each one of them contains the complete process from the ball's serve to dead, and it can be taken as an event unit.

Semantic units are automatically detected in these clips, and are considered independent each other.

The test data set has 32 clips, totally about 90 minutes. They are taken out from a real soccer game between Cameroon and Germany in FIFA World Cup in 2002. In the 32 clips, 17 of them are shoot events, 12 of them are card events. They are also segmented manually.

Table 1. Performance of Event Detection

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

Table 1 presents the performance of event detection of our method in terms of the two commonly used evaluation criteria: precision and recall. For shoot events, 14 of 17 are detected correctly, 3 of them are missing, no false detection. In the 3 missing clips, the shoots are all not very excited, 2 of them have no player close-up view, and one has no SMR. For card events, 10 of 12 are successfully detected, 2 of them are missing, no false results. In the 2 missing detection clips, 1 of them has no close&caption unit, another has no SMR unit.

The result is satisfactory. One reason is that the testing video clips are elaborately segmented in advance. Each of them is a complete event (interesting or uninteresting) segment grabbed from original video stream. With these data, the difficulty of event detection reduces very much. In fact, automatic event based video segmentation is difficult.

In the result, some events are missing if one semantic unit is absent. The possible reason is that we do not find out all useful evidences of events. The defined units are all so important that events strong depend on each of them.

Table 2. Performance of Semantic units Detection

S-U*	Truth	Detect	False	Miss
SMR	26	26	0	0
Aude	4	4	0	0
Goal	27	26	1	1
Capt	16	16	0	0
Close	53	47	10	6
CoCp*	13	13	1	0

*S-U – Semantic Units; CoCp – Close&caption.

Table 2 displays the performance of semantic units detection. We can see that the performance of SMR unit detection, Aude unit detection and Capt unit detection are quite perfect. The close-up unit detection is relative good. The reliability and validity of evidences in intermediate layer guarantees perfect performance of event detection.

6. CONCLUSIONS

In this paper, a semantic unit based event detection scheme is proposed. The scheme can be characterized as a three-layer framework. At the lowest layer, low-level visual features including color, texture, edge and motion are considered. At the intermediate layer, semantic unit concept is put forward and utilized to bridge the gap between low-level features and high-level semantics. A semantic unit can be taken as an intermediate semantic descriptor for video segment. At the highest layer, a Bayesian network based probabilistic framework is used to event inference with the observations of semantic units. We apply this scheme to shoot and card events detection in soccer videos. The experimental results demonstrate the validity and effectiveness of this method.

One contribution of this scheme is that it handles semantic units to bridge the semantic gap between low-level features and high-level events. The selection of semantic units is domain- and task-dependent. In the implementation, this scheme fully takes the domain-specific knowledge and video editing rules into consideration. The probabilistic reasoning involves statistics, machine learning, etc. It should reveal the casual relationship from observations to decisions, and temporal transition among these observations.

In the implementation in this paper, we only consider the casual relationship between observations and decisions, but ignore their temporal characteristics. The assumption that they are independent each other is reluctant. HMM and DBN will be considered in the future.

At present, the training and test data are grabbed from videos manually. Each of them is a bout that contains a complete event (interested and uninterested). In the future, automatic video segmentation based on event should be studied.

7. ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China (Grant No. 60135020 and 60121302).

8. REFERENCES

[1] M.R. Naphade, T.S.Huang, "Semantic Video Iindexing Using a Probabilistic Framework", *Proc. Int'l Conferecne on Pattern Recognition*, vol.3, Sep. 2000, pp. 83-88.

[2] B. Li, J. Errico, "Bridging the Semantic Gap in Sports", *Proc. IS&T/SPIE Conf. Storage and Retrieval for Median Databases*, vol. 5021, 2003, pp. 314-326.

[3] L. Xie, S-F, Chang, A. Divakaran, H. Sun, "Structure Analysis of Soccer Video with Hidden Markov Models", Proc. *IEEE Int'l Conf. on Acoustics, Speech and Signal Processing*, vol. 4, May 2002, pp. 4096-4099.

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

[5] N. Vasconcelos, A. Lippman, "A Bayesian Framework for Semantic Content Characterization', *Proc. Computer Vision and Pattern Recognition*, 1998, pp. 566-571.

[6] R. Qian, N. Hearing, and I. Sezan, "A computational approach to semantic event detection", *Proc. Computer Vision and Pattern Recognition*, June 1999, pp. 200-206.

[7] J. Terrillo, M. Shirazi, H. Fukamachi and S. Akamatsu, "Comperative Performance of Different Skin Detection of Human Faces in Color Images", *Proc. of IEEE Int'l Conf. Automatic Face and Gesture Recognition*, France, March, 2000.

[8] H.Pan,P.Beek, and M. Sezen, "Detection of Slow-motion Replay Segments in Sports Video for Highlights Generation", *Proc. IEEE Int'l Conf. On Acoustics, Speech, and Signal processing*,2001.

[9] X.F. Tong, Q.S.Liu, H.Q. Lu, and H.L. Jin, "Shot Classification in Sports Video", *Proc . Int'l Conf. On Signal Processing*, Aug. 31 – Sep. 4, 2004.

[10] C.Wolf, J.Jolin, and F.Chassaing, "Text Localization, Enhancement and Binarization in Multimedia Document", *Proc. Int'l Conf. Pattern Recognition*, vol. 2, Aug 2002, pp. 1037-1040.

[11] Cecil Huang, and Adnan Darwiche, "Inference in belief networks: a procedural guide", *Int'l Journal of Approximate Reasoning*, vol. 11, 1994,pp.1-45.