International Journal of u- and e- Service, Science and Technology
Vol.8, No. 12 (2015), pp.1-10
http://dx.doi.org/10.14257/ijunesst.2015.8.12.01

Research on Analysis of Sports Video based on the Statistics

Haitao Yang[1], Jia Wang[1] and Jingmeng Sun[2,*]

[1]Beijing University of Technology, Beijing 100000, china
[2]Physical Education Department, Harbin Engineering University,
Harbin 150001, China
yanghaitao@bjut.edu.cn

Abstract

We turn to statistical-based methods and propose a statistical inference approach by using Dynamic Bayesian Network, which is able to learn automatically from data set. By soccer video analysis is as an example, the proposed method is verified by experiment. We extract the color, shape and other low-level features from the video, to detect and identify 5 kinds of high-level semantic events using dynamic Bayesian network model. The experimental results show that our method is effective.

Keywords*: Video Analysis, Sports Video, Semantic Event Detection, Dynamic Bayesian Network*

1. Introduction

Grammar-based sport video analytical method is the one based on rules [1-2], easy for implementation and application. But it requires human setting rules or grammars for these reasons: firstly, in different competitions and photographic environments, it requires experts to set different rules as per experience, increasing difficulty in directly setting rules; next, in some sport videos [3-4], the relationship among events is not determined. It is uncertain and probabilistic association [5]. To overcome it, we'll discuss the method based on statistics. Unlike the method based on rules, the proposed method has some learning and adaption ability. Also it can take advantage of probabilistic relevance among events to improve effectiveness of event detection [6].

In previous work, some literatures investigated sport video analysis methods based on statistics. The most commonly used are Bayes Network (BN) and Hidden Markov Model (HMM), for instance, [7] used Bayes network to classify frame images in soccer videos to several typical scenes. In [8], the author introduced a sport event detection method based on HMM. However they have limitations for video analytics. Bayes network can perform quite well in classification. But it's a static classification model without capability to use fully contexts which change along with the time. HMM fits for processing time signals, like voice signal. But in video content analysis, its communication ability is restricted, primarily because video is a kind of multi-dimensional signal with both spatial information and temporal information. In light of all previous work, we introduce a more powerful sequence signal statistical tool, *i.e.*, dynamic Bayes network (DBN) [11-12], to analyze sport video contents. The new method, on one thing, extends the modeling ability of Bayes network to sequence signals by considering transition probability at each moment; on the other thing, it allows to use a few state variables at a similar time point, rather than only one state variable used by HMM. Based on them, we think dynamic Bayes network is more suitable to analyze sport video contents, especially semantic events and interrelationship among them [13-14].

[*] Corresponding Author

ISSN: 2005-4246 IJUNESST

2. Event Detection based on Dynamic Bayes Network

As for the detection of semantic events in video contents, it’s necessary to build effective mapping relationship between low-level features and high-level semantics. Here we utilize dynamic Bayes network to create the mapping. We’ll introduce how to set up dynamic Bayes network model as per domain knowledge, how to fetch effective low-level features and how to learn and infer high-level semantics.

2.1. Domain Modeling

We take football videos for example to analyze the five events: shot, corner ball, free kick, progress, suspension. They all are defined by game rules according to human understandings. They have rich high-level semantics. Apparently, it’s very difficult and inefficient to map directly from lower features like texture, shape and color to higher semantics. To avoid inefficiency, we suggest transforming it to an inference question, *i.e.*, higher semantic events consisting of lower elements, which can be mapped to low-level features. In the indirect mapping mode, effective mapping will form between lower features and elements. Then by statistical inference of the relationship formed among lower elements, we can detect and recognize high-level elements. Compared with direct mapping, we think indirect mapping is more useful.

Based on the idea, as well as game rules and television relay regulations, we determine the five scenarios in football videos as lower elements. As shown in Figure 1, they are respectively: (1) close-up shot and out-of-field shot; (2) medium shot; (3) midfield; (4) front court; and (5) penalty area. Close-up and out-of-field shots are pictures of people above the waist and audiences out of shooting site. Medium shots are pictures of one or some players in the field. Shooting scenes of midfield, front court and penalty area are long shots for different areas. In other sport videos, it’s required to redefine semantic events and scene elements according to characteristics of the game.

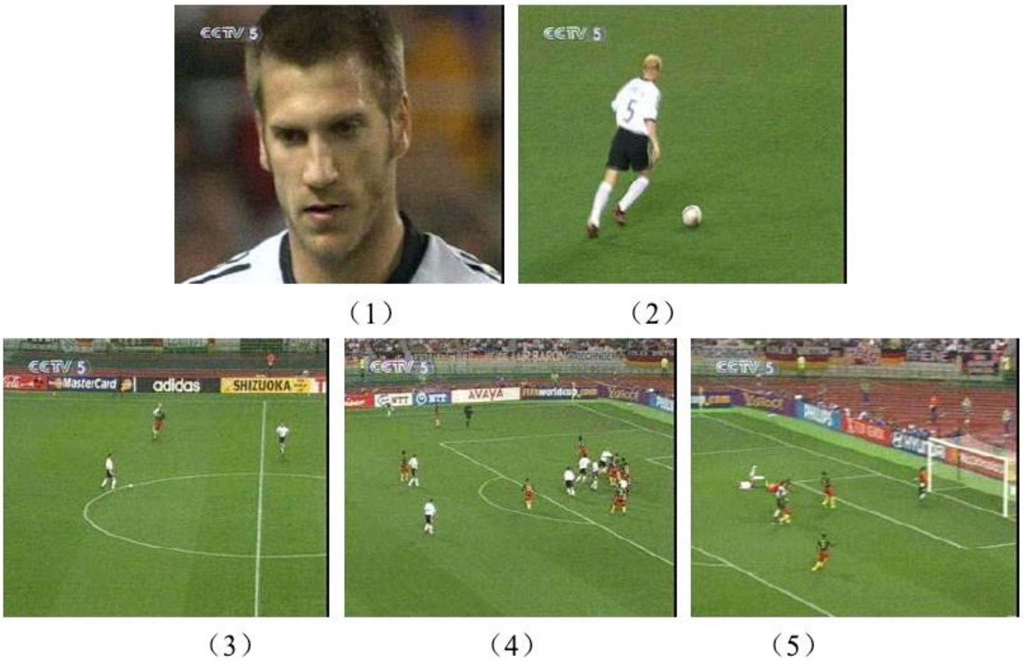

Figure 1. The Basic Scene in Soccer Video

Between these scene picture and semantic event are close contact, so according to these scenes context and transition probability can be judged events. A typical corner event shows some corner area of scene. Although the football video game content is different, but in order to facilitate the audience understanding of the game. The shooting mode are consistent.

Based on the above analysis, we construct such as dynamic Bayesian network model in Figure 2. The model has three layers, from top to bottom are event layer, element layer and observation layer. Through three layer network, the underlying features as statistical reasoning mode image to high-level semantic events.

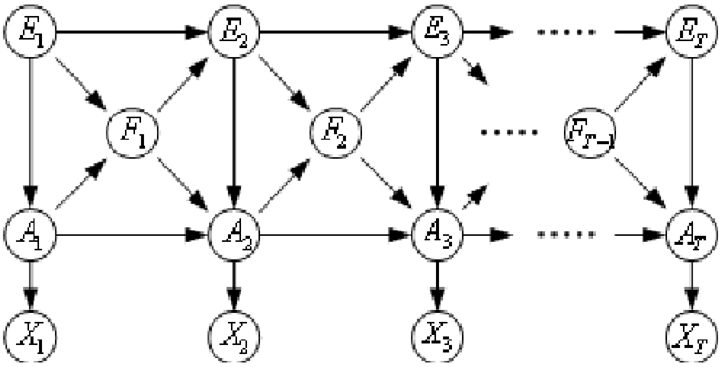

Figure 2. Dynamic Bayesian Network Model of Sports Video Analysis

2.2. Extraction of Features

In consideration of features of football videos, frame image features we shall pick up include the four descriptors: field area, player size, penalty area and midfield. With them, we can distinguish basic scenes in the videos.

2.2.1. The Field Area. The field area means the coverage of the game site in frame images, defined as follows:

$$f_{field} = \frac{n_{field}}{n} \tag{1}$$

Where, n_{field} is the number of pixels in the color range of the field in the image; n is total pixels of the image. We describe principal color detection algorithm in the following. The method can adaptively decide the current main color based on played video contents.

(1) Initially, the algorithm selects randomly K frame images in the middle section to put in the buffer queue;

(2) To eliminate impacts by illumination and shadows, cached images' color space should be converted from RGB to HSV space; then select H component to calculate their histogram $h(i)$; set i the chromaticity H with most pixels; $h(i) \geq h(j), j \neq i$, the initial main color range is [i-r, i+r], where r is radius of the main color;

(3) In the initial main color range, make iterative adjustments according to mean values to determine more accurate main color range. The process is as follows:

Calculated the average main color interval:

$$m = \frac{\sum_{k=i-r}^{i+r} kh(k)}{\sum_{k=i-r}^{i+r} h(k)} \tag{2}$$

Reset the main color range is [m-r, m+r].

Repeat the above steps until the main color interval does not change, or the number of iterations exceeds the threshold.

(4) Considered the process in the game, the main video color may change with time. It need adjust of the main color interval according to the current video content.

2.2.2. The Size of the Athletes. By binaryzation original frame images according to ground color, we can get binary images as in Fig. 3(b). The black area is field; white blocks are players. The biggest white block area is used as descriptor of player size:

$$f_{player} = \max(\frac{n_{region}}{n}) \tag{3}$$

In which, n is total pixels of the images; n_{region} is white area coverage in the field, which is calculated by the equation:

(1) Scan images in line; if the beginning and end of pixels in one line are continuous non-zero values, meaning it is out-of-field area; the algorithm skips without processing these pixels; for the rest pixels, it processes from the left to the right; if pixel P is a non-zero value, meaning it's a white area within the field; the algorithm determines to fix a tag as per surrounding pixels;

(2)After the above step, all areas are marked. But some areas are tagged differently due to conflicts. Scan again the whole image; re-mark images according to equivalent tag table; modify conflicted tags to minimal equal tags. Sum up the frequency of every tag's occurrence; the tag appearing the most frequently stands for the area with largest coverage in the field.

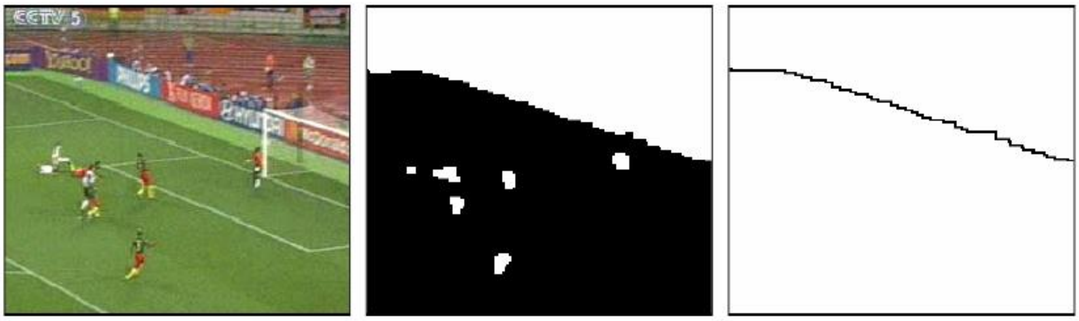
(a) the original image (b) two value image (c) the site outline

Figure 3. Forbidden Zone Detection

2.2.3. The Forbidden Zone. For the detection of penalty area, we suggest the method based on field contour. The football pitch has base lines and side lines (Figure 3(c)). When the penalty area appears in the image, base lines become longer and foul lines are shortening. With those observations, we define the descriptor of it like:

$$f_{gold} = \frac{n_{end}}{\sqrt{w^2 + h^2}}(1 - \frac{n_{side}}{w}) \tag{4}$$

Were w and h refers to width and height of frame images; n_{end} and n_{side} is the number of dots in base lines and sidelines. The method for detecting sidelines and base lines is described as follows:

(1) Clear away foul lines and bottom lines which interfere the detection, we remove the white area within the field from the binary image; then, with the use of Sobel operator, we carry out edge extraction to get profile of the field;

(2) On the basis of edge graph got previously, we utilize Hough Conversion with angle and position constraints to check out both side and bottom lines.

$$S_x = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & 0 \end{bmatrix} \quad S_y = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{5}$$

Hough transformation is for detecting straight lines in images. We get image center as origin, then any straight line in images can be expressed in polar coordinates as:

$$\rho = x\cos(\theta) + y\sin(\theta) \tag{6}$$

Where, (ρ, θ) defines a vector of the nearest point of origin. This vector and the linear are vertical. Finally, the linear statistical of the most votes is as detected line. Number of votes obtained by the straight line. Considering the bottom line and edge features, we add the corresponding angle and position constraints in the algorithm, remove some not meet the conditions of the candidate (ρ, θ), in order to improve the detection accuracy and reduce computational complexity.

The constraint condition of the bottom line:

$$\begin{aligned} &\theta \in [20,75] \cup [105,160] \\ &p \in [0.5r, r] \end{aligned} \tag{7}$$

Where r is the half image diagonal.

Constraint condition of sideline

$$\begin{aligned} &\theta \in [85,95] \\ &p \in [0.5h, h] \end{aligned} \tag{8}$$

Where, h is the height of the image.

2.2.4. Midfield. Descriptors representing the midfield are defined as:

$$f_{mid} = \frac{n_{mid}}{h} \tag{9}$$

Where, n_{mid} is the number of pixels in midcourt line. Center lines are detected by the same algorithm for side/bottom lines. What's different is the detection of midlines is to make edge extraction and Hough conversion of gray images of original frames. The point is when implementing Hough transformation, we consider only dots within the field, exclusive of those outside it.

2.3. Learning and Reasoning

Given dynamic Bayesian networks model before the structure, it have two tasks to complete in detecting and identifying the characteristic sequence of events. The first task is to learn, that is to estimate the probability distribution of the parameters in the network. After the estimated parameters from the training data set, the remaining task is reasoning, that is to calculate the maximum probability of the sequence of events in feature observation sequences.

According to the structure is known and nodes are hidden, dynamic Bayesian network learning has different methods. In this paper, we established the dynamic Bayesian network model is known structure and node can be observed. According to the statistics of training samples directly estimates the conditional probability distribution between nodes. Dynamic Bayesian network model can be expressed as the following probability distribution function in Figure 2.

2.3.1. The Observation Layer. We use Gaussian Mixture Model to represent the observation probability distribution function of the underlying Element:

$$P(x_t \mid A_t) = \sum_{m=1}^{M} w_m N(x_t, u_m, \sigma_m) \tag{10}$$

Gauss mixture model is trained using Expectation Maximum algorithm.

Set N sample x_k is observation vector of the element in the training data, training specific steps is the iterative process as follows.

(1) According to the parameters of the existing mixed Gauss model, calculated the sample belongs to the probability of each sub distribution:

$$P(m \mid x_k) = \frac{N(x_k, u_m, \sigma_m)}{\sum_{m=1}^{M} w_m N(x_k, u_m, \sigma_m)} \tag{11}$$

(2) Re-estimation of parameters of mixed Gauss model:

$$\begin{aligned} w_m &= \frac{1}{N}\sum_{k=1}^{N} P(m \mid x_k) \\ u_m &= \frac{\sum_{k=1}^{N} P(m \mid x_k) x_k}{\sum_{k=1}^{N} P(m \mid x_k)} \\ \sigma_m &= \frac{\sum_{k=1}^{N} P(m \mid x_k)(x_k - u_m)(x_k - u_m)^T}{\sum_{k=1}^{N} P(m \mid x_k)} \end{aligned} \tag{12}$$

(3) Repeat the above process, until Probability $\sum_{k=1}^{N} \ln P(x_k)$ of the mixture Gauss model does not increase obviously.

2.3.2. The Element Layer. The conditional probability distribution of nodes

$$\begin{aligned} P(A_t = \sigma \mid A_{t-1} = \sigma', E_t = \omega, F_{t-1} = 0) &= A_\omega(\sigma, \sigma') \\ P(A_t = \sigma \mid A_{t-1} = \sigma', E_t = \omega, F_{t-1} = 0) &= \pi_\omega(\sigma) \end{aligned} \tag{13}$$

Where, Switch instructs F_t to open distribution of the condition probability

$$P(F_t = 1 \mid A_t = \sigma, E_t = \omega) = A_\omega(\sigma, end) \tag{14}$$

2.3.3. The Event Layer. Event nodes related to conditional probability distribution

$$\begin{aligned} & P(E_1 = \omega) = \pi(\omega) \\ & P(E_1 = \omega \mid E_{t-1} = \omega', F_{t-1} = f) = \begin{cases} \delta(\omega', \omega) & f = 0 \\ A(\omega', \omega) & f = 1 \end{cases} \end{aligned} \tag{15}$$

Where, π said probability distribution of sequence events, A said the transfer probability distribution of events.

3. Experimental Analysis and Results

The experiment has two sections. Frist of all, we evaluate the event detection effect of the method; next, considering users' requirements for automatic extraction of wonderful fragments, we regard shot, corner kick, free kick as the same fabulous fractions to assess the performance of the proposed algorithm. The testing and training data are collected from videos of four different football matches, which held in different sites and were broadcasted by different TV companies. We chose totally 54 video clips lasting from a few to over ten minutes. Before the experiment, we annotated manually all events in them to use as real reference data. Those clips constitute a video data set which lasts more than two hours. For that reason, we used half of them as training set and the rest as testing set.

Table 1. Experimental Results of Event Detection

Event	Correct	Error	Missing	Precision	Recall
Corner	25	16	2	60%	90%
Free kick	15	7	6	68%	70%
Shooting	41	16	10	71%	80%
The game proceed	116	15	11	88%	91%
Interruption of game	55	13	11	80%	84%
Total	253	68	40	78%	86%

Table 2. Experimental Results of Highlight Extraction

	Correct	Error	Missing	Precision	Recall
Frame	50060	10500	6540	82%	88%
Fragment	100	30	2	83%	98%

Table 1 shows experimental results of event detection. The precision rate is on average 78% and the mean recall ratio reaches 86%. The algorithm achieved higher accuracy rate of detection in the progress and suspension of matches. In view of big content changes of them, the results are satisfactory. But it didn't do well in detecting corner ball and free kick, maybe because there's great similarity in the formation of elements of the two events. To enhance the performance, it needs finer elements and more effective algorithm for feature extraction and element recognition.

Table 2 gives experimental results of extracting wonderful clips by our method. Those fragments relate to shot, corner kick and free kick which may lead to goals. It lists evaluation results based on frames and fragments. Frame-based assessment is to calculate if every frame is classified correctly. The latter concerns if extracted highlights are correct and if some lost. Compared with event detection, these results look much better as the experiment loosened the requirement regarding classification of events. Although more than six thousand frames were lost, only two wonderful clips were not detected. That means most frames were lost in the anterior and posterior boundary of wonderful clips which have been detected. It is acceptable. In fact for human beings, it's very hard to define the exact boundary of one wonderful clip.

Based on the automatic analysis algorithm for football videos, we realized a football video management prototype system (Figure 4). When you open football videos, you can choose automatic analyzing tools to detect and discern automatically wonderful events in videos. Results are shown in the event list on the right side of the screen. Besides, it provides a “Tags on” view. You can modify manually annotations of events there or compute statistically events in video.

4. Conclusion

In this paper, we propose sports video analysis based on Dynamic Bayesian network statistical method. Firstly, we established a multilayer dynamic Bayesian network model based on sports video domain knowledge, including the observation layer, element layer and the event layer. In this model, the high-level semantic events is component of base, and then mapped to low-level features, so as to avoid the high-level semantics and low-level features mapping difficulties. Secondly, we introduce the learning and inference algorithms based on statistics. Finally, based on these algorithms, dynamic Bayesian network model can learn knowledge from the training samples, and then carries on the inference between semantic events in sports video.

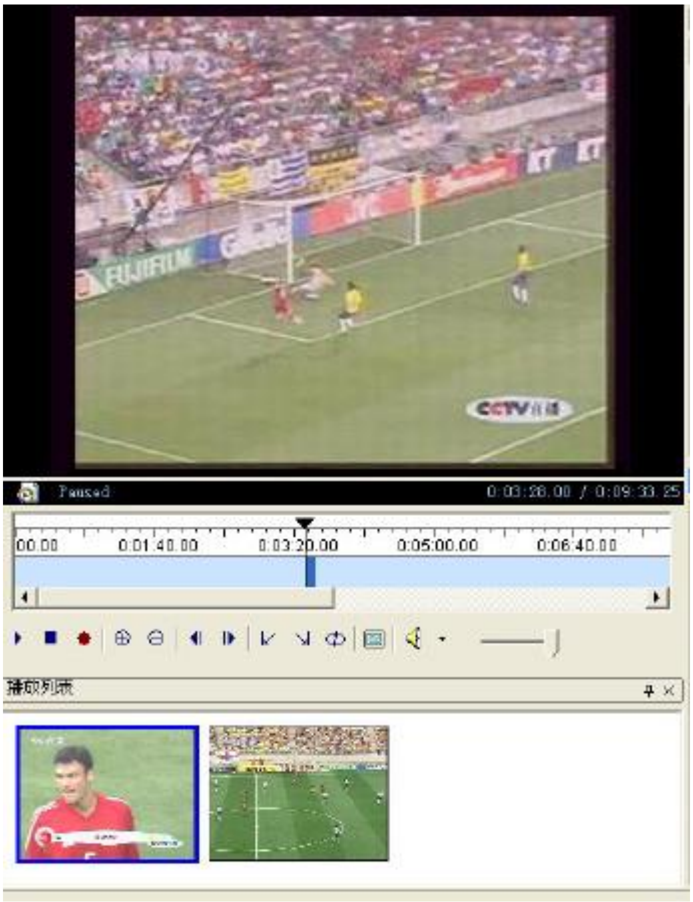

Figure 4. Soccer Video Management System

Acknowledgment

This work was supported by The Fundamental Research Funds for the Central Universities. No. HEUCF151601.

References

[1] L. Wang, “And realize multi-mode teaching video semantic analysis”, Nanjing University of Science and Technology, **(2014)**.

[2] K. Liu, “Research and design feasibility basketball video tactical real-time tracking and analysis system”, Capital Institute of Physical Education, **(2014)**.

[3] W. Zhang, “Application Research on the technology of broadcast sports video based on Android platform, North China University of Technology, **(2014)**.

[4] D. L. Wei, “Study on detection method for soccer video highlights model based on HCRF”, Xi'an Electronic and Science University, **(2014)**.

[5] H. Han, “Research on tennis serve video event detection”, Inner Mongolia Normal University, **(2014)**.

[6] P. Wu, X. Q. Lin, H. T. Li, Z. Gao and X. S. Zhan, “Sports video classification based on color texture and SVM”, Journal of Fujian Normal University (Natural Science Edition), vol. 2, **(2014)**, pp. 34-41.

[7] “The probability distribution PI said sequence starting events”, A said the distribution of transfer probability between events.

[8] M. Luo, Y. F. Ma and H. J. Zhang, “Pyramidwise Structuring for Soccer Highlight Extraction”, Proceedings of IEEE Pacific-Rim Conference on Multimedia, **(2013)**.

[9] L. Y. Duan, M. Xu, X. D. Yu and Q. Tian, "A Unified Framework for Semantic Shot Classification in Sports Videos", Proceedings of ACM International Conference on Multimedia, **(2002).**

[10] Y. X. Cui, "Two dimensional body joint characteristics of sports video annotation based on Computer Engineering", vol. 4, **(2014)**, pp. 252-257.

[11] G. Xu, Y. F. Ma, H. J. Zhang and S. Q. Yang, "A HMM Based Semantic Analysis Framework for Sports Game Event Detection", Proceedings of IEEE International Conference on Image Processing, **(2013)**.

[12] K. P. Murphy, "Dynamic Bayesian Network: Representation, Inference and Learning", PhD Dissertation, University of California, Berkeley, **(2012)**.

[13] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceeding of the IEEE, **(1989)**.

[14] L. Y. Duan, M. Xu, T. S. Chua, Q. Tian and C. S. Xu, "A Mid-level Representation Framework for Semantic Sports Video Analysis", Proceedings of ACM International Conference on Multimedia, **(2013).**

Author

Haitao Yang, he received his B.S degree from Capital Institute of Physical Education, and received his M.S degree from Hebei Normal University. He is a lecturer in Beijing University of Technology. His research interests include Physical education and training.

.