International Journal of Database Theory and Application
Vol.8, No.6 (2015), pp.313-322
http://dx.doi.org/10.14257/ijdta.2015.8.6.28

Research on Analysis of Sports Video Multi-Pattern Fusion

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

[1]*Beijing University of Technology, Beijing 100000, china*
[2]*Physical Education Department, Harbin Engineering University, Harbin 150000, china*
tennis_jia@aliyun.com

Abstract

In order to effectively integrate multimodal information and multilayer constraints, we present a unified probabilistic framework for sports video analysis. Based the framework, three instances of statistical models are constructed and compared. Experimental results indicate our method with multimodal fusion processes semantic events in sports video more effectively.

Keywords: *Video Analysis, Sports Video, Semantic Event Detection, Highlight, Multimodal*

1. Introduction

With the method based on statistics, we'll discuss further sport video content analysis method fusing multimode information. Semantic events in sport videos are in essence multimodal. In television relay, multimode information is integrative used to present video contents like subtitles, narrator's voices, on-site sounds, camera movements, scenarios and images *etc*. It is incomplete to analyze only one mode. For more effective analysis of events, it's required to study the analytical method which fuses multiple patterns. On the other hand, semantic events in those videos are not isolated. There's some logical or consequential relationship among them. In previous paper, we discussed event detection and recognition with the use of the contextual relationship based on dynamic Bayes network. Now on that basis, we'll explore how to fuse multimode information, which is a key issue we're facing here [1-2].

In recent years, the fusion of multimode information has become a hot topic in the field of sport video analysis [3-5]. Firstly we introduce the related work. Most of the multimode fusion analysis methods mentioned in previous literatures considered the detection of an isolated event. Unlike them, we propose to detect many events and analyze comprehensively the association among them.

Multi-level analysis methods based on statistics are built on the probabilistic graphical model, such as Hidden Markov Model (HMM), dynamic Bayes network (DBN) and their variants. By combining visual graphical model representation and effective reasoning and learning methods, such solutions made fairly good effects. Xie [6-7] *et al*. applied hierarchical HMM for unsupervised clustering to discover layered structure of video contents. Differently, we introduced the method based on dynamic Bayes network model to do the same work. Through learning of training samples, we fulfilled the detection and recognition of wonderful events in the football match [8].

Based on DBN, we developed a common sport video analysis framework. Compared with previous work mentioned above, our approach can not only integrate multimode information for event detection and also deal with hierarchical relationship among events. Although the two functions were stated in previous papers, they as a whole were not paid

* Corresponding Author

ISSN: 2005-4270 IJDTA

too much attention to. Now with a universal probabilistic framework and being putting together for mutual complement, it's possible to realize robust and effective event detection. Based on the framework, three multimode multi-level analytical models were generated for sport videos, including Factorial Hierarchical Hidden Markov Model （FHHMM）. Coupled Hierarchical Hidden Markov Model (CHHMM) and Product Hierarchical Hidden Markov Model (PHHMM). In order to compare their performance, we experimented on "in-play" and "out-of-play" events in football match videos, by means of the three models and traditional hierarchical HMM. Results showed that Product Hierarchical Hidden Markov Model (HHMM) gained the best effects.

2. Multimode Multi-Level Semantic Analysis Framework of Sport Videos

For the semantic analytics of sport videos, multimode information and multi-level constraints are important foundations to do that. But in existing papers, there isn't a solution based on statistics, which can combine them together. Here we proposed a multimode multi-level analysis framework on the basis of DBN. Then based on it, we designed three models such as: FHHMM, CHHMM and PHHMM. Firstly we'll give expression form of them based on DBN; then, we'll discuss the learning and reasoning algorithm.

2.1. Expression of Models

Dynamic Bayes network is one kind of directed probabilistic graphical model. Its parameter can be put as (Λ, Θ). The first group Λ refers to DBN's structure-related parameters inclusive of node quantity in each frame, network topological structure *etc*.; the second group Θ stands for the conditional probabilistic distribution to which all connection lines in network relate, node's initial probability *etc*.

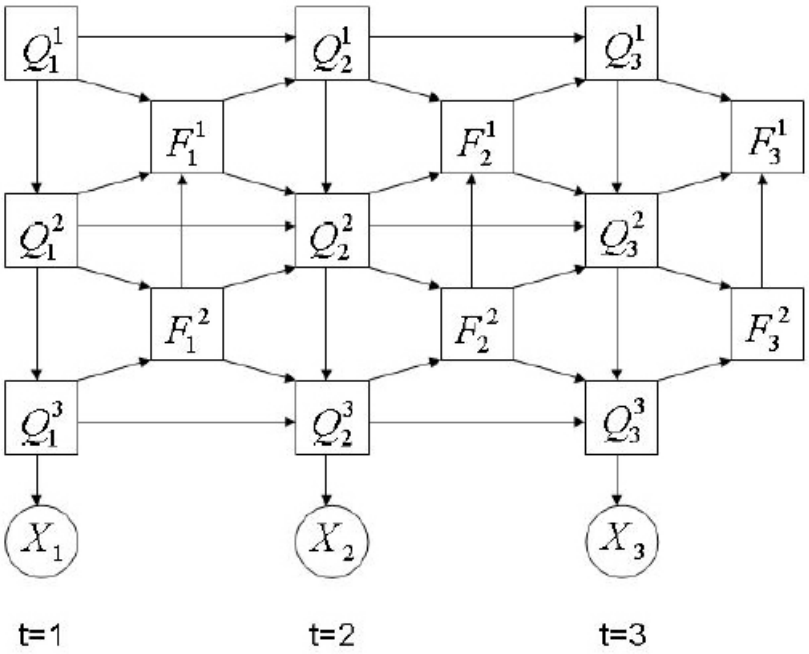


Figure 1. HHMM DBN of a Three Layer

Figure 1 is the graphical structure of a three-level HHMM represented as DBN. In it, Q_t^d means state variable at time point t in the level d; the total number of states in d is n_d; F_t^d is indicator variable; F_t^d=1 if and only if Q_t^d's sublevel HMM *i.e.* Q_t^{d+1} moves with time to the end state. What should be noted is if $F_t^d = 1$, then in every level below, there exists $F_t^{d'} = 1(d' > d)$.

After viewing the above structure, we'll give definitions of all related probabilistic distributions. Since the local topological structure of nodes in different levels is varied, their respective probabilistic distribution function differs.

The top floor:

$$P(Q_1^1 = i) = \pi_1^1(i)$$

$$P(Q_t^1 = j \mid Q_{t-1}^1 = i, F_{t-1}^1 = f) = \begin{cases} \delta(i,j) & \text{if } f = 0 \\ a_1^1(i,j) & \text{if } f = 1 \end{cases} \tag{1}$$

The middle layer:

$$P(Q_1^d = i \mid Q_1^{d-1} = k) = \pi_k^d(i)$$

$$P(Q_t^d = j \mid Q_{t-1}^d = i, Q_t^{d-1} = k, F_{t-1}^d = f, F_{t-1}^{d-1} = g) = \begin{cases} \delta(i,j) & \text{if } f = 0 \\ a_k^d(i,j) & \text{if } f = 1 \text{ and } g = 0 \\ \pi_k^d(j) & \text{if } f = 1 \text{ and } g = 1 \end{cases} \tag{2}$$

$$P(F_t^d = 1 \mid Q_t^d = k, Q_t^{d+1} = i, F_t^{d+1} = f) = \begin{cases} 0 & \text{if } f = 0 \\ a_k^d(i, end) & \text{if } f = 1 \end{cases} \tag{3}$$

The bottom:

$$P(Q_t^D = j \mid Q_{t-1}^D = i, Q_t^{D-1} = k, F_{t-1}^{D-1} = 0) = a_k^D(i,j)$$

$$P(Q_t^D = j \mid Q_{t-1}^D = i, Q_t^{D-1} = k, F_{t-1}^{D-1} = 1) = \pi_k^D(j)$$

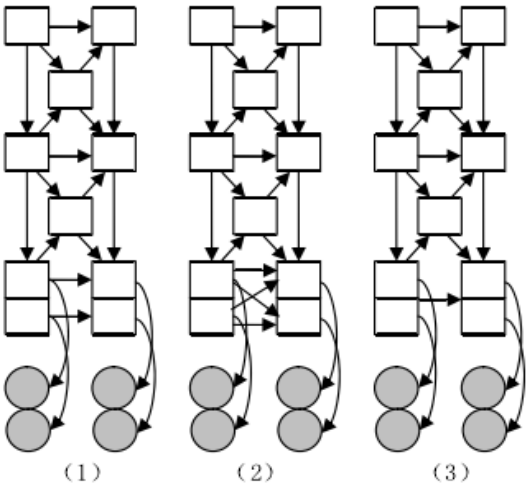
$$P(F_t^{D-1} = 1 \mid Q_t^{D-1} = k, Q_t^D = i) = a_k^d(i, end)$$

$$P(X_t \mid Q_t^D = i) = N(X_t, u_i, \sigma_i) \tag{4}$$

Where, π_k^d represents the starting state probability distribution of state $Q_t^{d-1} = k$. $a_k^d(i,j)$ is the state transition probability, $N(X_t, u_i, \sigma_i)$ observed probability characteristics of X_t as the bottom status is i. For simplicity, in this paper we use the multivariate Gauss model to represent the observation probability. In fact, you can also use other forms such as hybrid probability distribution, Gauss distribution.

Multi-mode fusion processing is initially got researchers' attention in the field of speech recognition. Known as Coupled Hidden Markov Model (CHMM) dynamic Bayesian networks have been applied widely

From this idea, we on the traditional HHMM were extended, proposed the sports video multi-mode multi-level analysis framework. Figure 2 shows the fusion of different multi-model strategy based on this framework. Diagram of the box said state variables, and circles represent were from two different patterns of the observation vector.



(1) FHHMM; (2) CHHMM; (3) PHHMM

Figure 2. Dynamic Bayesian Network Model for Multi-Mode Multi-Level

2.2. Learning and Reasoning

Given the expression of above models, we can do learning and reasoning to analyze video contents. Learning means to predict parameters of models by according to training samples which were manually marked. Reasoning is to give observation sequences and determine status sequences when the probability is the biggest.

In our models, observation sequences fetch multi-pattern feature data from videos; status sequences correspond to semantic events and temporal relations among them.

Compared with learning, the solution of reasoning problem is simpler. So we resolve it firstly. To start, we convert multi-level DBN model to HMM and then get solution with Viterbi algorithm [9]. It is an optional reasoning method and easier for implementation. Considering DBN is in nature a Bayesian network, we can perform reasoning based on Bayes network Junction Tree algorithm [10].

In order to apply Viterbi algorithm for inference, we transform the proposed multi-level DBN to the following HMM:

$$\pi(\mathbf{i}) = \pi_1^1(i_1)\prod_{d=2}^{D}\pi_{i_{d-1}}^{d}(i_d) \tag{5}$$

$$a(\mathbf{j}\mid\mathbf{i}) = a_{i_{r-1}}^{r}(i_r, j_r)\prod_{d=r+1}^{D}a_{i_{d-1}}^{d}(i_d, end)\pi_{j_{d-1}}^{d}(j_d) \tag{6}$$

$$b(X_t\mid\mathbf{i}) = N(X_t, \mu_{i_D}, \sigma_{i_D}) \tag{7}$$

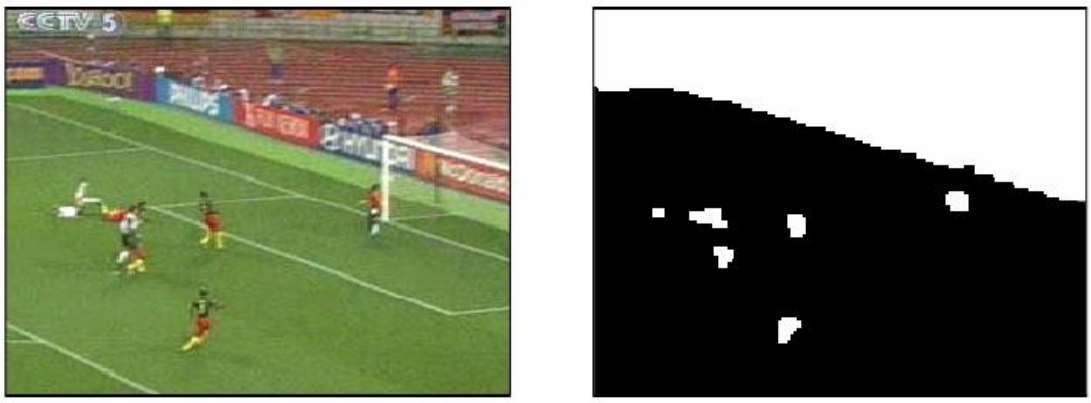
Where, State $i=[i_1,...,i_D]$ is the state vector of the multilayer DBN model, then, the total number of state Cain Markov model is $N=i_1,i_2...i_D$. Set the length of the sequence is T, using Viterbi algorithm to solve optimal state sequence.

3. Detection of In-Play and Out-Of-Play Events in Football Videos

To verify and compare the performance of DBN model, we applied the above models to detect "in-play" and "out-of-play" events in football videos. "In-play" events are defined as the time for the football game goes on normally. "Out-of-play" events, also called dead ball, refer to what happens when the ball passes on the ground or in the air wholly over goal line or sidelines; or when the game is called by a judge. The detection of those events is very significant to automatically generate video abstracts and make higher semantic analysis. To be specific, by detecting game progress and suspension, we can remove suspended video fragments to generate more simplified video abstracts.

Firstly we extract from video streams the color and movement features to use as observation data in different patterns. As stated in [11], those features originate from different modes, among which the correlation is very little. But at the same time, for the detection of "in-play" and "out-of-play" events, the information of mutual supplement in the two modes are valuably referential. Therefore, it requires a method which can fuse effectively modal information and utilize fully contextual restraints to detect events. Here we didn't use audio features. That's because they're not quite distinguishable for the detection of "in-play" and "out-of-play" events after our observation.

[12] Suggested detecting the game is ongoing or paused by according to shot classification and some heuristic rules. The method required manually setting rules and determining threshold as per experience. In [6], Xie *et al*. used exercise intensity and main color ratio as characteristics, with two groups of HMM to stand respectively for "in-play" and "out-of-play" events. Then based on HMM's output probability, they segmented and recognized events by dynamic programming. On that basis, Xie developed a new method based on HHMM, which was confirmed effective by the experiment on detecting "in-play" and "out-of-play" events. Here we also implemented a method based on HHMM for event detection. We used it as standard for comparison.



(a) The Original Image (b) The Binary Image

Figure 3. Feature Extraction Based on Dominant Color

In our experiment, we f*etc*hed color properties based on the main color. As the pitch is a major part of most scenarios, the main color in football videos stands for the occurrence of the pitch. For every frame image, we can get a binary image as seen in Figure 3(2) by using the main color to binary it. We use geometric moment to represent shape features of the pitch thereof. Set image size $M \times N$, the moment of order (p,q) is defined like:

$$m_{pq} = \sum_{y=0}^{N-1} \sum_{x=0}^{M-1} x^p y^q f(x,y) \tag{8}$$

Based on these moments, we calculate the area respectively (Equation 9), the level of variance (Equation 10) and the vertical variance (Equation 11), three groups described as color feature:

$$A = \frac{m_{00}}{MN} \tag{9}$$

$$\sigma_x = \frac{1}{M}\sqrt{m_{20}/m_{00} - m_{10}^2/m_{00}^2} \tag{10}$$

$$\sigma_y = \frac{1}{N}\sqrt{m_{02}/m_{00} - m_{01}^2/m_{00}^2} \tag{11}$$

In order to extract the motion features, first, we calculated for each motion vector of images by block matching, then calculate the exercise intensity (Equation 12) and entropy (Equation 13):

$$m_i = \frac{1}{K}\sum\sqrt{v_x^2 + v_y^2} \tag{12}$$

$$m_e = -\sum_{i=0}^{8}\frac{h(i)}{K} Log \frac{h(i)}{K} \tag{13}$$

K is the total number of the motion vector $v = [v_x, v_y]$, $h(i)$ is the number of motion vectors in the eight directions. $i = 0$ indicates a motion vector of static. Motion feature includes color features for other helpful information of event detection.

4. Experimental Analysis and Results

The experiment has two objectives:

(a) to evaluate whether our proposed multi-level DBN models like FHHMM, CHHMM and PHHMM can effectively realize fusion processing of multimode information;

(b) to compare the performance of the proposed method and others.

To fulfill the first goal, we carried out a system based on traditional HHMM as benchmark. The system uses one-mode information for processing, in other words, it applies only the color features for training and recognition. Then, it utilizes only motion features for the same task. The system contains higher and lower nodes. Nodes in higher levels correspond to “in-play” and “out-of-play” events; nodes in lower levels correspond to hidden elements. For the second goal, we implemented a system based on feature fusion as benchmark. Despite using still HHMM, this system combines motion features and color features together as observation input. It uses feature fusion method instead of decision fusion method. Unlike HHMM-based system, the proposed FHHMM, CHHMM and PHHMM create separately elements and observation for different modes. The number of different element’s states affects the function of those models. In our experiment, we find whether the state number is too less (<5) or too more (>9), our models’ performance is reduced. In this case, we choose the best result of each model for comparison.

The data set for testing includes 20 video clips which last from a few minutes to over ten minutes. They were chosen from five sessions of football match videos, MPEG-1 format, size 352x288, 25 frames per second. At every 0.5s, we f*etc*h from video stream the color and motion features to decrease calculated amount. The “in-play” and “out-of-play” events in testing data set were manually remarked as real data beforehand. Cross-Validation experiment was conducted to train and assess those models. That is to say,

every time 90% data is used for training and the rest for testing. Repeat ten times till all data are tested in turn.

Regarding video analysis results acquired by different models, we evaluate them based on video frames and segments. Frame-based evaluation is made to compare true data with results got by automatic analytics by frame, calculating the percentage of right marked frames versus the total. The evaluation can reflect the sensitivity of each model to local changes (Table1). There HHMM_C refers to the mere use of color features; HHMM_M refers to the mere use of motion features. Other models use the two features simultaneously.

Table 1. The Experimental Results Based on Frame

Model	HHMM_C	HHMM_M	HHMM	FHHMM	CHHMM	PHHMM
Accuracy rate	78.6	63.7	77.8	81.4	86.3	84.6

From Table1, we see PHHMM achieved the highest rate of accuracy. Compared with models using only single mode, the three multimode multi-layer DBN models obtained better results. They also performed better than HHMM using feature fusion mode. In our experiment, the feature fusion method was even worse than models utilizing only color features. The reason may lie in too much interference in motion features. The combined application of color and motion features actually degraded the effectiveness of those features. Conversely, the proposed fusion model proved better adaptability.

The assessment based on fragments (Table 2) is to compare analysis results with truthful data by according to the fetched event fragments. When detected fragments overlap temporally those in real data, it's believed the detected events are correct. If several detected event fragments overlap one similar real event, the detection results except the first are all thought detected falsely. In order to get an entire appraisal of analytical results, apart from commonly used recall rate R and precision ration P, we introduced the harmonic mean F -value for the purpose. F -value is defined as:

$$F = \frac{2RP}{(R+P)} \tag{14}$$

Table 2. The Experimental Results Based on the Fragment

Model	HHMM_C	HHMM_M	HHMM	FHHMM	CHHMM	PHHMM
Precision	87.6	50.6	69.8	36.4	36.3	73.6
Recall	71.6	84.7	84.8	100	99.3	90.6
F-value	78.6	65.7	76.8	53.4	54.3	80.6

In Table 2, except PHHMM, FHHMM and CHHMM led to higher recall rate but lower precision ratio. The reason for that is over-segmentation in their output results, *i.e*. lots of short-lasting event fragments occurring there. Hence a few event fragments would match with the same one real event. According to the evaluation algorithm, except the first, all other event clips are considered wrong, which thus resulted in lower precision rate. The over-segmentation incurred perhaps because of excessive consideration of local changes due to weaker contextual global limitations. Differently, the preliminary experiment indicated PHHMM didn't have the problem and achieved satisfactory effect. It suggested that PHHMM can not only make effective use of dynamic interaction among information in different modes but also ensure multi-level constrained relationship of contexts. It's a promising semantic analysis model of videos.

5. Conclusion

In this paper, we proposed a dynamic Bayesian network based on a fusion of multi model information and multi-level constraint sports video analysis framework. On the one hand, multi-level analysis based on dynamic Bayesian network can express the domain knowledge for the topological structure of intuitive. On the other hand, the learning and inference algorithms can effectively establish a statistical interaction between multi-pattern information.

In order to evaluate the performance of the model, we take the game of soccer video detection interrupt event as an example to test. As a control, we also implement the traditional HHMM method. The experimental results show that these methods can better comprehensive multi pattern information and multi-level constraint. PHHMM achieved the best results. Compared with the single mode HHMM, both experimental results based on the frame and the experimental results based on fragmentary, performance increases about 15%.

Acknowledgement

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

References

[1] M. Han, W. Hua, W. Xu and Y. H. Gong, "An Integrated Baseball Digest System Using Maximum Entropy Method", Proceedings of ACM International Conference on Multimedia, **(2012)**.
[2] Y. Wang, Z. Liu and J. C. Huang, "Multimedia Content Analysis Using Both Audio and Video Clues", IEEE Signal Processing Magazine, **(2010)**.
[3] M. Barnard, J. M. Odobez and S. Bengio, "Multi-Modal Audio-Visual Event Recognition for Football Analysis", Proceedings of IEEE Workshop on Neural Networks for Signal Processing, **(2013)**.
[4] M. Xu, L. Y. Duan, C. Xu and Q. Tian, "A Fusion Scheme of Visual and Auditory Modalities for Event Detection in Sports Video", Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, April **(2013)**.
[5] M. Petkovic, V. Mihajlovic, W. Jonker and S. D. Kajan, "Multi-Modal Extraction of Highlights from TV Formula 1 Programs", Proceedings of IEEE International Conference on Multimedia and Expo, **(2012)**.
[6] L. Xie, S. F. Chang, A. Divakaran and H. Sun, "Unsupervised Discovery of Multilevel Statistical Video Structures Using Hierarchical Hidden Markov Models", Proceedings of IEEE International Conference on Multimedia and Expo, **(2008)**.
[7] L. Xie, S. F. Chang, A. Divakaran and H. Sun, "Feature Selection for Unsupervised Discovery of Statistical Temporal Structures in Video", IEEE International Conference on Image Processing, Barcelona, Spain, September **(2013)**.
[8] F. Wang, Y. F. Ma, H. J. Zhang and J. T. Li, "Dynamic Bayesian Network Based Event Detection for Soccer Highlight Extraction", Proceedings of IEEE International Conference on Image Processing, Singapore, October **(2004)**.
[9] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceeding of the IEEE, vol. 77, no. 2, **(1989)**.
[10] T. A. Stephenson, "An Introduction to Bayesian Network Theory and Usage", IDIAP Research Report, February **(2000)**.
[11] 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, November **(2003)**.
[12] L. Xie, S. F. Chang, A. Divakaran and H. Sun, "Structure Analysis of Soccer Video with Hidden Markov Models", Proceedings of International Conference on Acoustic, Speech, and Signal Processing, May **(2002)**.

Author

Jia Wang, She received her B.S degree from Beijing Sport University, and received her M.S degree from Beijing Sport University. She is a lecturer in Beijing University of Technology. Her research interests include Physical education and training.