01 Jan 2003

A HMM Based Semantic Analysis Framework for Sports Game Event Detection

Gu Xu
Missouri University of Science and Technology

Yu-Fei Ma

Hong-Jiang Zhang

Shiqiang Yang

A HMM BASED SEMANTIC ANALYSIS FRAMEWORK FOR SPORTS GAME EVENT DETECTION

Gu Xu[†], *Yu-Fei Ma*[‡], *Hong-Jiang Zhang*[‡], *Shiqiang Yang*[†]

[†]Dept. of Computer Science, Tsinghua University, Beijing, China (100084)
[‡]Microsoft Research Asia, 5/F Sigma Center, 49 Zhichun Road, China (100080)

ABSTRACT

Video events detection or recognition is one of important tasks in semantic understanding of video content. Sports game video should be considered as a rule-based sequential signal. Therefore, it is reasonable to model sports events using Hidden Markov Models. In this paper, we present a generic, scalable and multi-layer framework based on HMMs, called SG-HMMs (*Sports Game HMMs*), for sports game event detection. At the bottom layer of this framework, event HMMs output basic hypotheses based on low-level features. The upper layers are composed of composition HMMs, which add constraints on those hypotheses of the lower layer. Instead of isolated event recognition, the hypotheses at different layers are optimized in a bottom-up manner and the optimal semantics are determined by top-down process. The experimental results on basketball and volleyball videos have demonstrated the effectiveness of the proposed framework for sports game analysis.

1. INTRODUCTION

Efficient searching and retrieving video content have become more and more difficult due to the ever increasing of video data. Semantic analysis is a natural way for video management and accessing. However it remains a challenging issue. In recent years, so many efforts have been made on event detection, especially for sports videos. Based on the assumption that excited speech always occurs right after baseball highlights, R. Yong [1] *et al* extracted highlights by excited speech detection. In [2], camera motion was extracted directly from MPEG data to detect basketball events. Though not all important events were recognized, it showed the effectiveness of motion features in semantic analysis. The semantic object extraction was used in [3] and [4] for sports game analysis. In [3], several high-level events in tennis video were detected by reasoning under the court-line and players' location information. [4] focused on soccer games, in which court-line, player and ball information were extracted for identifying the camera location of the court. Court-line and plays' position are important for sports game analysis. However object extraction algorithms are usually not reliable. In [5], a soccer video analysis system was presented to classify video sections into three camera views and obtain play/break status of the game by rules. The work in [6] combined the statistical and semantic object model for sports classification, in which color histogram clustering was employed first, and verified by semantic object information. A different approach was presented in [7], in which rules are trained by an entropy-based inductive tree-learning algorithm. However, video context constraints were not utilized yet. Hidden Markov Models (HMMs) are kinds of temporal trainable models which have been successfully used in speech recognition [8]. Recently, HMMs were also applied to video content analysis. In most of work, such as [9], HMMs were used to model temporal inference, which were often manually built in terms of topology. On the other hand, HMMs were utilized to recognize isolated events, for example, the strokes in tennis [10].

In this paper, we propose a generic multi-layer HMM based framework for sports game semantic analysis, called *Sports Game HMMs* (SG-HMMs). The different layers represent different semantic levels. The higher the layer is, the higher the level of semantic it represents is. We suppose that sports game video is a sequence of basic events with a set of rules. Each event corresponds to a pre-trained HMM and the rules are modeled by multi-layer composition HMMs. Consequently, the optimal recognition result is the event sequence which gives the maximum likelihood. With this framework, the semantic events are segmented and recognized simultaneously. We applied the proposed framework to basketball and volleyball games semantic analysis. The encouraging results have been obtained in our experiments.

The rest of this paper is organized as follows. In Section 2, we describe the framework of *Sports Game HMMs* (SG-HMMs) in details. In Section 3, we apply this framework to basketball and volleyball games analysis. Experimental results are given in Section 4. Finally, Section 5 concludes this paper.

2. SPORTS GAME HMMS

There are three key issues in video semantic analysis. That is, 1. how to utilize the prior knowledge in semantic analysis; 2. how to resolve the conflict between recognition and temporal segmentation; 3. how to sufficiently utilize context information. Although the three issues have been addressed in previous works, they were seldom considered together. In this paper, we proposed a multi-layer framework for semantic analysis in sports video, in which the three key issues are all well considered.

We assume that video sequence is composed of basic events and the relationship between events follows a set of rules. We also assume that the basic events are numerable in a specific domain, which makes it possible to create a model for each basic event. Therefore, Hidden Markov Models (HMMs) are selected as the kernel of the proposed framework. In this framework, each basic event is modeled by a pre-trained HMM,

0-7803-7750-8/03/$17.00 ©2003 IEEE

and the composition rules are mapped to hidden markov process. The basic event HMMs form the bottom layer and the upper layers are composed of composition HMMs.

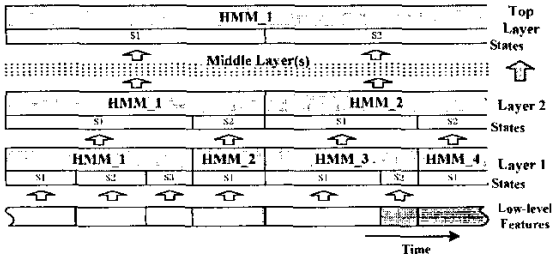

Fig. 1 Framework of Multi-level SG-HMMs

The proposed multi-layer framework introduces the context constraints at different semantic levels. As shown in Figure 1, the low-level features are directly fed into the bottom layer to generate the hypotheses of basic events. It is reasonable to assume the causality between basic events follows the first order Markov process. Therefore, a composition HMMs layer is added above bottom layer to introduce constraints of higher level semantics. In this manner, we may hierarchically add composition layers into the framework elevating semantic level. The top layer only has one HMM, whose optimal output determines the final recognition in all layers. The other layers are composed of more than one HMM. The state number of HMMs at same layer is equal, which is determined by

$$N_k^s = \begin{cases} N_{k-1}^{HMM} & k > 1 \\ c & k = 1 \end{cases} \quad (1)$$

where N_k^s is the state number at k^{th} layer and N_{k-1}^{HMM} is the number of HMMs at $(k\text{-}1)^{th}$ layer. The constant c is defined manually at the bottom layer.

If let A denote the state transition probability distribution, B denote the observation probability distribution in states, and π be the initial state distribution, HMMs are defined as follow:

$$\lambda = (A, B, \pi) \quad (2)$$

In (2), the specification of state number and observation are omitted. As mentioned above, in our framework, there are two kinds of HMMs corresponding to event recognition and the presentation of composition rules. The HMMs for basic event recognition are standard HMMs, in which the Baum-Welch algorithm is used to re-estimate the probability measure A and B. In contrast, a different way is adopted in composition HMMs to calculate the probability measures. In composition HMMs, each state corresponds to a HMM at the layer directly below it. Therefore, measure B at this layer is taken by the likelihoods outputted from the corresponding HMMs. Measure A is regarded as posterior probability. Thus, the transition probability from event x to y is equal to the conditional probability of event y given event x, $P(y|x)$. According to the Bayesian equation, probability measure A is computed by

$$A_{x,y} = P(y|x) = \frac{P(xy)}{P(x)} = \frac{N(x,y)}{\sum_{k \in E} N(x,k)} \quad (3)$$

where, $A_{x,y}$ denotes the element in matrix A in row x of column y, and E denotes the set of HMMs in the lower layer. $N(x,y)$ is the number of occurrences of the adjacent pair of event x and y (y follows x). In our implementation, transition probability matrix A is calculated by taking the counts of event pairs according to manually annotated event transcriptions. The probability measure π is regarded as a part of measure A if we add a virtual "begin" state into HMM, say, the first state of the HMM state sequence.

In our framework, prior knowledge and context information are respectively modeled by two types of HMMs. Meanwhile, the conflict between segmentation and recognition is solved. Specifically, given a video clip, all possible HMM sequences guided by the composition HMM are aligned on the whole clip. Finally, event boundaries are determined by the boundaries of HMMs in the optimal HMM sequence. In fact, the HMMs at bottom layer may be substituted by other models, such as TDNNs.

3. APPLICATIONS

The proposed framework is a generic solution to sports video analysis. When an application domain is given, only observation feature and the number of semantic layer are required to be determined in advance. System automatically learns the relationships between semantics from the training data. Currently, we applied the proposed semantic analysis framework, (SG-HMMs) to basketball and volleyball games, as shown in Figure 2.

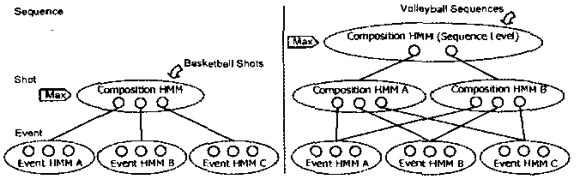

Fig. 2 SG-HMMs in Basketball and Volleyball

Although basketball and volleyball game have different characters and game rules, they have their own distinctive motion patterns which provide significant cues. In [11], we have proposed a motion based representation, including three motion curves corresponding to horizontal, vertical and radial motions respectively. Based on the three motion curves and their differential curves, a 6-dimension vector is obtained. In this work, we still employ this 6-dimension vector as observation vector in SG-HMM.

As the elements of observation vector are real, the continuous HMMs are used at bottom layer. Single Gaussian is adopted for VQ in current applications. In our implementation, we employ a complete connected six-state HMM as a general prototype for all basic events in both basketball and volleyball games. The experimental results show that the six-state HMM is reasonable. Additionally, increasing the number of HMM states has no distinct improvement in recognition performance.

3.1. Application to basketball games

According to game rules, recoding manners and viewers' interests, we defined 16 events for basketball games. They include : 1. offence at the left court; 2. offence at the right court; 3. fast break to the left court; 4. fast break to the right court; 5.

lay-up at the left court; 6. lay-up at the right court; 7. shot at the left court; 8. shot at the right court; 9. track player to the left; 10. track player to the right; 11. lay-up in the close-up view; 12. shot in the close-up view; 13. foul shot in the close-up view; 14. general close-up; 15. wipe; 16. stillness. Some meaningless clips, such as pure camera zoom-in and zoom-out without any object motions, are categorized into stillness. Moreover, the general close-up event also accepts some meaningless clips. In this manner, the probability of the model breakdown is effectively decreased.

We employ two-layer SG-HMMs in basketball game analysis, which is the basic form of the proposed framework. In this form, the event recognition/segmentation is carried out at the shot level. That is, only the relationships between events within one shot are modeled. Therefore, in this application, basketball videos are segmented manually into shots at first, including training and testing sequences. For training sequences, we further segment each shot into pre-defined events. Each annotated event label is individually used to train bottom-layer HMMs, while event transcriptions of shots are utilized to compute probability measures in composition HMMs. During recognition process, each shot of testing sequences is fed into SG-HMM respectively. The SG-HMM maximizes the probability for each shot independently and cut it into pre-defined events automatically.

3.2. Application to volleyball games

The recording manners in volleyball games are more regular than that in basketball games. Therefore volleyball videos often show clearer context-dependence, especially in the shot level, namely, the relationship between shots. With such observations, we extend the basic form used in basketball application to three-layer SG-HMM for utilizing the new context constraints at a higher semantic level. The new HMM layer is added for shot relationship modeling. In such three-layer SG-HMM, the event recognition at shot level is carried out at the lower two layers, and the top layer models the semantic structure of the entire video sequence. In this manner, the event segmentation, recognition and shot classification are completed in this three-later SG-HMM.

As discussed above, the SG-HMM in volleyball spans over two semantic levels, that is, shot category level and basic events level. Therefore, the semantics in volleyball are pre-defined in two levels respectively. The shot categories in this work contain specific semantic information, such as serving team. With such semantic information of shots, more accurate events may be pre-defined, which cannot be recognized within a shot. The pre-defined events in volleyball are divided into 14 categories. Some events are further extended to a number of events by attaching such high level semantic attributes as misplay and score. The 14 event categories include 1. serve at the right or left court; 2. offence at the right or left court; 3. attack to the right or left court; 4. stroke at the right or left court; 5. block at the right or left court; 6. zoom-in when playing; 7. jump serve/serve in close-up view by right or left team. 8. cheer in the right or left court; 9. track player to right or left; 10. general close-up; 11. stillness; 12. wipe; 13. slow-motion of stroking in right or left court; 14. slow-motion in close-up view.

Similar to the basketball applications, all training and testing videos are segmented manually into shots. However, in the training sequences, each shot is not only segmented into basic events, but also classified into a shot category manually. The training process on bottom-layer HMMs is consistent, while the upper-layer HMMs are trained differently. In basketball application, all shots are considered same. But for volleyball, in order to explore shot relationships, shots are classified into several categories. Therefore, in the second layer of SG-HMM, the number of HMMs equals the number of shot categories, instead of one. The probability measures of each second layer HMM is calculated by event transcriptions with the same shot category label. The top layer still contains only one HMM, that is to say, all volleyball sequences are considered as same. Consequently, we build the top layer HMM by category transcriptions in the sequence level.

Though the three-layer SG-HMM is able to accept whole video sequences directly, we still provide the shot information during recognition process in order to reduce the computational complexity. The recognition process is divided into two steps, namely, shot level recognition and sequence level recognition. In shot level recognition, each shot is parsed by every two-layer shot model, which provides basic hypotheses both on shot category and basic event. In sequence level recognition, those hypotheses are optimized under the constraints of relationships between shots. It seems that the two steps are corresponding to event detection and shot classification respectively, however, they are interactive and tightly connected by probabilities.

4. EXPERIMENTS

We evaluate the performance of the applications in basketball and volleyball games respectively. Considering the ambiguity of semantics, we carry out evaluation experiments by user study. A distributed evaluation system is designed for user assessment from network. The subjects are invited and required to give scores to computer generated results online. In our experiment system, the event information is extracted from the database and presented to users by speeches and texts simultaneously when an event is emerging. At each end of shots, the subjects are required to select an assessment: good, neutral or bad. If a subject thinks the current shot cannot be described by our predefined events, the subject is allowed to refuse the assignment. We define the rate of users' satisfaction as follow:

$$Rate_{sat} = \frac{N_{Good} \times 100 + N_{Neutral} \times 50}{N_{Good} + N_{Neutral} + N_{Bad}} \times 100\% \qquad (4)$$

where N_{Good} is the number of the "good" assessments, $N_{Moderate}$ is the number of the "neutral" assessments and N_{Bad} is the number of the "bad" assessments.

4.1. Basketball events detection

In this experiment, the total duration of testing videos is more than 6 hours. As training data, about 20 sample clips are used for training each basic event HMM and 380 event transcriptions of shots for building the top layer HMM. 15 subjects are involved in the experiment.

The results of user study are listed in Table 1. It indicates that the events we defined are reasonably complete, as only

1.35% assignments were refused by subjects. The average user satisfaction rate approaches 75%, which indicates that our framework is successfully employed in basketball games.

Table 1 Basketball Evaluation

Shot Num	Assessment				Score
	Good	Neutral	Bad	Refu.	
85	576	156	112	33	77.5%
441	3385	918	685	102	77.1%
532	3822	1093	769	97	76.9%
699	3201	885	834	50	74.1%
470	2990	758	919	30	72.2%
359	1500	321	457	9	72.9%
2586	**15474**	**4131**	**3776**	**321**	**75.0%**

4.2. Volleyball events detection

In the experiment of volleyball, two sessions, about 40 minutes, are used for training and seven sessions, about 110 minutes, for testing. Because the occurrences of different events are extremely uneven, we still cannot obtain adequate samples for some events though 40-minute videos are employed. However, most of events are well trained by at least 10 samples. More than 400 transcriptions of shots are utilized for building upper-layer HMMs. Totally, 10 subjects are invited for assessment.

The assessments are given in Table 2. Only 0.5% of assignments are refused by subjects, which are much fewer than that in basketball. It is because we have extracted more semantics by adopting new constraints at the shot level. The average user satisfaction exceeds 76%, which is also higher than that in basketball. Usually, the strict defined events result in lower satisfaction rate for more recognition errors. Although the events defined in volleyball are stricter than that in basketball, a higher score is obtained. Therefore, it shows that the third layer constraints effectively improve the recognition results.

Table 2 Volleyball Evaluation

Shot Num	Assessment				Score
	Good	Neutral	Bad	Refu.	
203	275	464	1074	30	72.0%
218	276	484	1312	4	75.0%
107	125	209	727	2	78.4%
157	138	286	1138	17	82.0%
260	409	494	1682	6	74.6%
217	273	403	1487	4	78.1%
135	187	241	889	1	76.7%
1297	**1683**	**2581**	**8309**	**64**	**76.4%**

5. CONCLUSIONS

In this paper, we have presented a novel framework for sports game video semantic analysis, called SG-HMM. The proposed framework is extendable in architecture, which can introduce context constraints and give outputs at different semantic levels. With the assumption that the sports events follow Markov process, two kinds of HMMs are defined in our framework. The event HMMs compose the bottom layer of SG-HMM, which produce the hypotheses based on low-level features. The composition HMMs in the upper layers optimize the hypotheses and pick out the optimal event combinations by context information. Simultaneous temporal segmentation and recognition also is the distinctive feature of our framework. By applying this framework to basketball and volleyball games, we have demonstrated its effectiveness in semantic analysis for sports games. By providing different training data, our framework can be applied to other sports games easily.

6. REFERENCES

[1] Yong Rui, Anoop Gupta, Alex Acero, "Automatically Extracting Highlights for TV Baseball Programs", Proc. ACM Multimedia, Los Angeles USA, pp. 105-115, Oct. 2000

[2] Y.-P. Tan, D. D. Saur, S. R. Kulkarni and P. J. Ramadge, "Rapid estimation of camera motion from compressed video with application to video annotation", Circuits and Systems for Video Technology, IEEE Transactions, 2000

[3] G. Sudhir, J.C.M. Lee, A.K. Jain, "Automatic classification of tennis video for high-level content-based retrieval", Proc. Content-Based Access of Image and Video Database, IEEE International Workshop on, pp. 81-90, 1998.

[4] Y.H. Gong, Lim Teck Sin, Chua Hock Chuan, Hongjiang Zhang, Masao Sakauchi, "Automatic parsing of TV soccer programs", Multimedia Computing and Systems, Proceedings of the International Conference on, pp. 167-174, 1995.

[5] P. Xu, L. Xie and S.F. Chang, et al, "Algorithms and Systems for Segmentation and Structure Analysis in Soccer Video", IEEE Int. Conf. on Multimedia and Expo, Tokyo, Japan, Aug. 2001.

[6] D. Zhong, S.F. Chang, "Structure Analysis of Sports Video Using Domain Models", IEEE Int. Conf. on Multimedia and Expo, Tokyo, Japan, Aug. 2001.

[7] W. Zhou, A. Vellaikal and C. C. Jay Kuo, "Rule-based video classification system for basketball video indexing", Proceedings on ACM multimedia workshops, 2000.

[8] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of the IEEE, 1989.

[9] B. Li, M. Ibrahim Sezan, "Event Detection and Summarization in Sports Video", IEEE Workshop on Content-Based Access of Image and Video Database, December 2001.

[10] M. Petkovic, Z. Zivkovic and W. Jonker, "Recognizing Strokes in Tennis Videos Using Hidden Markov Models", IASTED Int. Conf. Visualization, Imaging and Image Processing, Marbella, Spain, September 2001.

[11] G. Xu, Y.F. Ma, H.J. Zhang and S.Q. Yang, "Motion Based Event Recognition Using HMM", IEEE Int. Conf. on Pattern Recognition, 2002.