01 Jan 2002

Motion Based Event Recognition Using HMM

Gu Xu
Missouri University of Science and Technology

Yu-Fei Ma

Hong-Jiang Zhang

Shiqiang Yang

Motion Based Event Recognition Using HMM

Gu Xu[†], Yu-Fei Ma[‡], Hong-Jiang Zhang[‡], Shiqiang Yang[†]
[†]Dept. of Computer Science, Tsinghua University, Beijing, China (100084)
[‡]Microsoft Research, Asia. 5F Sigma Center, 49 Zhichun Road, China (100080)

Abstract

Motion is an important cue for video understanding and widely used in many semantic video analysis. In this paper, we present a new motion representation scheme in which motions in a video is represented by the responses of frames to a set of motion filters. Each of these filters is designed to be most responsive to a type of dominant motion. Then we employ Hidden Markov Models (HMMs) to characterize the motion patterns based on these features and thus classify basketball video into 16 events. The evaluation by human satisfaction rate to classification result is 75%, demonstrating effectiveness of the proposed approach to recognizing semantic events in video.

1. Introduction

Research efforts indexing and retrieval of video data have gone through three stages to address such requirements. The first stage focuses on video structure analysis to facilitate structured browsing. In the following stage, the focus is on similarity-based retrieval, influenced largely by content-based image retrieval (CBIR) research. However, similar to that in CBIR, the retrieval results using low-level visual features usually are far from satisfactory. This fact has motivated the effort to perform semantic analysis, the third stage, by which the semantic events will be extracted.

In recent years, there are many works in video semantic analysis, especially for sport videos. In [1], camera motion is extracted directly from MPEG data to detect basketball events. Though not all important events can be recognized, it shows the efficiency of motion features in semantic analysis. The work in [2] introduced a relatively constrained approach to extract court-line and players' location information in tennis video analysis. In [3], a soccer video analysis system was presented to classify video sections into three views and obtain play/break status of the game by rules. The system is robust to different conditions, but the semantic information extracted is not sufficient for normal users. A different approach, was presented in [4], in which rules are trained by an entropy-based inductive tree-learning algorithm. Although more important events can be detected, no context information of video is utilized.

Video is a context-sensitive media, so it is analyzed more effectively by sequential analysis tools. Hidden Markov Models (HMM) are kinds of temporal training models used successfuly in speech recognition [5], and they have been applied to the video content analysis in constrained conditions [6][7]. In [8], the strokes in tennis is detected by HMM. But the system lacks of robustness due to the difficulty of automatic object segmentation.

In this paper, we present a new approach to video event analysis. In this approach, motions between two video frames are represented by energy redistribution function. The temporal sequence of such energy redistribution functions derived from a video sequence is then filtered by a set of motion filters, each is designed to be most responsive to a type of dominant motion. Such a filter process converts a video sequence into a tempoal sequence of the filter responses in which distinct temproal patterns corresponding to high-level concepts are present. Then, we use HMM to segment shots of basketball video into semantic clips with 16 concepts of basketball game based on such temporal features. Compared with the related works, our approach can recognize more comprehansive events in basketball.

The rest of this paper is organized as follow. In Section 2, we describle how to extract sequencal features from video based on motion in details. The application to basketball video will be introduced in Section 3. Also, the using of HMM will be disscused in this section. The evaluation resutls by user study are given in Section 4. Lastly, Section 5 concludes this paper.

2. Feature Extraction

There are two key ideas in proposed new motion representation converting a video to a temporal feature sequence, which is discribed in detail in this section. First, motions between two frames of a video sequence are viewed as an energy redistribution process, which is represented by energy redistribution function. That is, the video sequence will be converted into a sequence of such functions. Second, a set of motion filters are designed, each is most responsive to a type of dominant motion. These filters are applied to the motion energy

1051-4651/02 $17.00 © 2002 IEEE

redistribtuon sequence of the video, producing a sequence of reponses to each filter that present the motion patterns in the video sequence.

2.1. Energy Redistribution

Our motion representation feature is drived from the motion vector fields (MVF) between video frames. We use the MVF determined by block-based motion estimation algorithms (BMA). Though the real motion often cannot be obtained by BMA, the lost is trivial compared to its efficiency, in particular when videos are in MPEG format where motion vector fields are readily available. Furthermore, our proposed representation views the motion vectors between two frames as the forces to alter the distribution of "energy" associated with each block, and measures the change between two frames by an energy redistribution function.

More specifically, each block in MVF is viewed as a basic energy container. We assume that all of blocks in the initial frame have the same amount of energy. Motion vectors are figured as the outside forces that cause energy exchange between blocks, as show in Fig. 1. Then, the change of energy distribution will exhibit motion features.

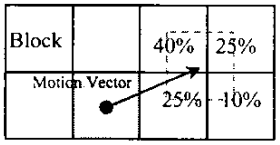


Figure 1. Energy redistribution

The redistribution of energy only depends on the position of correspondent block in the next frame. The energy at block(x, y) is denoted by $E_{x,y}$. So the energy distribution update is represented as

$$E'_{x,y} = \frac{\sum_{i,j} \left(overlapS_{i,j,x,y} \times E_{i,j} \right)}{W_b^2}, \quad i,j \in [1, W_b] \qquad (1)$$

where $overlapS_{i,j,x,y}$ denotes the overlap portion of the rectangular region corresponding to block(i, j) in previous frame and block(x, y) in current frame, and W_b is the size of blocks. If blocks move out of frame boundary, we place blocks just in frame by decreasing the magnitude of vectors to keep the amount of energy.

2.2. Weight Templates

The energy redistribution function only provides a way to represent motion between two frames. However, for video content understanding, we need convert such measure of energy redistribution into temporal motion patterns, hopefully each reflects a motion event quantitatively. For this purpose, we have designed a few motion filters, each of which is a weight matrix with the same dimensions of blocks as the video frames. Elements in a weight matrix are denoted by $w_{x,y}$. When we apply each of these filters to the energy redistribution field associated with a frame of a video, the energy response of the frame is defined by

$$E_R = \sum_{i,j} E_{i,j} \times w_{i,j}, \quad i,j \in [1, W_b] \qquad (2)$$

It is clear that by arranging the weights in a weight matrix with different values and orders, we can change the sensitivity of the filter to different motion forms. In other words, if we can design a set of weight matrixes, each is sensitive to a particular type of motions, we can apply them in detecting such motion patterns. This is the motivation of introducing the motion filters in our approach. This is clearly seen in Fig 2, where each plot contains three curves and the bold one is generated by the filters placed at the right. A crest on curves indicates the presence of a certain motion, the type and shape of crests show the direction and characters of the motions.

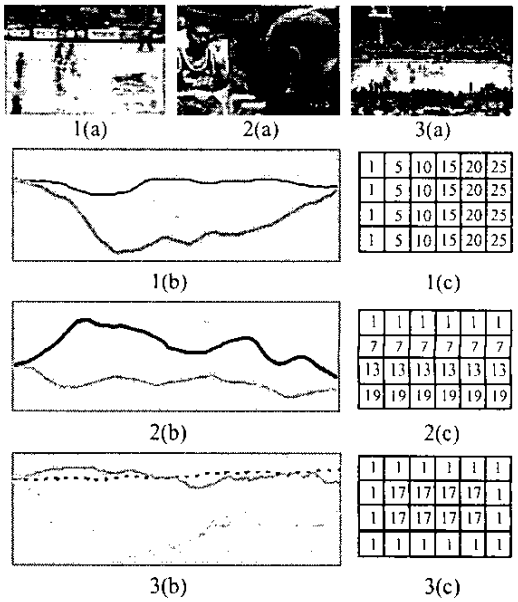


Figure 2. Examples of feature curves. (a) Key frame of clips, (b) Feature curve, (c) Example for weight template; 1. Horizontal motion, 2. Vertical motion, 3. Radial motion.

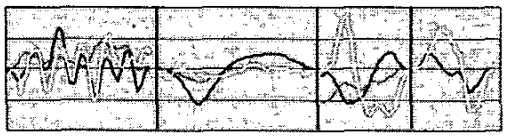
2.3. Sequential Feature Curves

To perform motion analysis with the proposed approach, we first apply the three filters shown in Fig.2 to a video sequence to produce a motion feature sequence of the video. Like in audio signal processing, a sliding window is used in calculating motion features of a video sequence. The first frame in the window is the initial

frame with the even energy distribution. Then, the energy redistribution function defined by (1) is calculated for each as more frames are added into the window one by one, till the last frame in the window reached. This process produces a sequence of energy redistribution images and each image is applied to each of the three motion filters, respectively, to produce a sequence of energy value, as defined by (2). Next, the mean of the energy value sequence generated by a filter is calculated and all the three means, each produces by a motion filter are combined into one sample vector to represent the motion of the frame sequence in the sample window. The three motion filters are designed to detect three kind of dominant motions: horizontal, vertical and radial, respectively, as show in Fig. 2. The width of the sliding window and the sampling frequency (defined by the number of skipped frames when the window slides) will determine the accuracy of results. So it will be easy to balance the computation complexity and the performance by adjusting those two parameters.

3. Basketball Game's Events Recognition

As an application of the proposed video representation and motion analysis scheme, we have developed a system to automatically detect events in live basketball videos using the representation and analysis scheme. Block-based motion fields of test videos are obtained using full-search algorithm. Then, three temporal feature sequences defined by (1) and (2) are produced It is observed that the correspondences between a number of semantic concepts and the temporal feature sequences are clearly presented, as shown in Fig 3. To establish such correspondence between all basketball events and observed feature patterns, thus, performing event extraction from video, Hidden Markov Models approach is used.



(a) (b) (c) (d)

Figure 3. Examples of concepts presented on curves. (a) Tracking, (b) Layup, (c) and (d) Two types of wipe in "BASKETBALL.mpg" video (in MPEG-7 video set)

HMMs are statistical models for sequential data. An effective application of HMM is in speech recognitions. In referring to recognition work in [9], we have developed a propose a new framework of using HMM in video content analysis. In this framework, shots in videos are viewed as sentences in speech and event clips as words. In this way, we represent video context at the semantic level. In our basketball analysis system, 16 conceptual categories (or events) of basketball games are defined: 1. team offence at left court; 2. team offence at right court; 3. fast break to left; 4. fast break to right; 5. lay-up at left court; 6. lay-up at right court; 7. shot at left court; 8. shot at right court; 9. tracking player to left; 10. tracking player to right; 11. lay-up in close-up view; 12. shot in close-up view; 13. foul shot in close-up view; 14. close-up; 15. wipe; 16. stillness. We consider these semantic events the minimal recognition unit and have trained a HHM model for each event. Here we request that events are self-contained in the generality of cases. In other words, shots in basketball video can be looked as sentences composed of those events. Then we build automatically the semantic net by training data to add knowledge rules in recognition. Finally, Viterbi algorithm is used to segment and recognize events in shots. In order to avoid over segmentation resulting in short segments meaningless for human understanding, we define several rules and apply them in the post-processing.

3.1. Training HMM

The HMM used in this paper is the continuous single Gaussian HMM. From a general view point, we define the topology of all initial models by the same prototype. Besides begin and exit states, the general structure is a complete connected six-state HMM. Since some events are distinct on original feature sequences and some on the difference sequences, we combine three original and three difference sequences into a six-dimension vector as the observation vector, input to HMM. The training samples are labelled manually.

3.2. Recognition by HMM

All of events are context dependent. In our framework, the relationship between semantic events is considered grammar rules in sentences. We have prepared shot transcriptions manually based on events defined and compute the transition probabilities by

$$p(i, j) = N(i, j) / N(i) \quad N(i) \neq 0 \qquad (3)$$

where $N(i)$ is the total number of occurrences of event i, $N(i,j)$ is the total number of c-occurrences of event i and j. The recognition is done on shot level and the best event transcription for each shot will be given by HMM. Viterbi algorithm segments shots by the maximum likelihood estimation and the transition probabilities represent a kind of the posterior probability. So, the product of the two probabilities will be the finial recognition probability. Finally, the concept transcription with the maximal recognition probability is regarded as the result.

4. Experimental Results

We have performed experimental evaluations of the our proposed approach on the basketball videos. The total duration of our experimental videos is more than 6 hours and over 2500 shots, including MPEG7 test video. All videos are first segmented in to shots manually. In our implementation, the width of sliding widow is 5 frames and sample period is one frame. We provide about 20 standard clips for each concept and 380 concept transcriptions of shots for creating semantic net.

To perform objective evaluations, a web-based evaluation system is developed for user study on internet. Users can evaluate our algorithm shot by shot online. When a shot is played, the events information can be obtained from database, which is presented to users by speech and text in real-time. When a user has finished watching a shot, he/she is asked to choose an assessment from three ranks: good, moderate and bad. If a user thinks the content in a shot beyond the pre-defined 16 concepts, he may refuse this assignment. We have invited 15 people to visit our web site, and evaluate out system and obtained 23705 assignment records. The rate of users' satisfaction is computed as

$$SatRate = \frac{N_{Good} \times 100 + N_{Moderate} \times 50}{N_{Good} + N_{Moderate} + N_{Bad}} \times 100\% \quad (4)$$

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

Shot Num	Assessment				Score
	Good	Mod.	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%**

Table 1. Experimental results

Shot Class	Shot Num	Score
Wide-angle view	841	61.8%
Close-up view	1511	80.9%
Wipe	99	92.4%

Table 2. Performance in different shot classes

The evaluation results are listed in Table 1. It is indicated that the 16 concepts defined are reasonably complete, since users have only refused 1.35% assignments. The average user satisfaction rate approaches 75%. In addition, the system has classified shots automatically into three classes: wide-angle, close-up and wipe. The classification performance for each of these classes is shown in Table.2. It is seen that wipe shots in basketball videos are distinguished by our system with the best result. Because of unapparent motion, our system can not always distinguish "layup" from "shot" and "left offence" from "right offence" successfully. So the score of field shots is the lowest.

5. Conclusion

In this paper, we have presented a new motion content representation scheme and its applications in basketball game video classification. The key contributions in the new scheme are a content representation based on the energy redistribution function and a set of motion filters to which the response of energy redistribution function present distinguishable motion patterns. Based on this representation, a new framework for video content analysis based on HMMs is proposed and applied to basketball video classifications. The experiments on about 6 hour's basketball game video proved the effectiveness of the proposed representation scheme and its application in video classifications.

Our feature extraction algorithm is a general method and can be applied to other sequential analysis tools. We are currently seeking ways to represent color and texture in temporal forms which can be used in our system.

References

[1] 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

[2] G. Sudhir, J. C. M. Lee and A. K. Jain, "Automatic Classification of Tennis Video for High-level Content-Based Retrieval", IEEE Multimedia, 1997.

[3] P. Xu, L. Xie and S.-F. Chang, et al, "Algorithms and Systems for Segmentation and Structure Analysis in Soccer Video", IEEE International Conference on Multimedia and Expo, 2001.

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

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

[6] C. Morimoto, Y. and L.S. Davis, "Recognition of Head Gestures Using Hidden Markov Models", International Conference on Pattern Recognition, 1996.

[7] T. Starner, "Visual Recognition of American Sign Language Using Hidden Markov Models, Master's Thesis", MIT Media Laboratory, February 1995.

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

[9] S. Young, et al, "The HTK Book (for HTK Version 3.0)", http://htk.eng.cam.ac.uk.