Sports Highlight Detection from Keyword Sequences Using HMM

Jinjun Wang[1, 2], Changsheng Xu[2], Engsiong Chng[1], Qi Tian[2]

[1]CeMNet, School of Computer Engineering, Nanyang Technological University, Singapore 639798
jjwang@pmail.ntu.edu.sg, aseschng@ntu.edu.sg
[2]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
{xucs,tian}@i2r.a-star.edu.sg

Abstract

Sports video highlight detection is a popular topic. A multi-layer sport event detection framework is described in this paper. In the mid-level of this framework visual and audio keywords are created from low-level features and the original video is converted into keyword sequence. In the high-level the temporal pattern of keyword sequences are analyzed by HMM classifier. The creation of visual and audio keyword can help to bridge the gap between low-level features and high-level semantic. The use of HMM classifier can automatically find the temporal change character of the event instead of rule based heuristically modeling to map certain keyword sequence into events. We experiment our model on soccer games and some promising results are obtained.

1. Introduction

Effective semantic classifications of video clips will help to solve the problem of managing and accessing huge amount of video data nowadays. Compared with other videos types such as news and movies, sports videos have well defined content structure and domain rules. A long sports game can be divided into parts and only a few of these parts contain certain highlights which are interesting to human. For example, in a soccer game, there are events such as goal, corner kick, free kick, etc. It is not difficult for human beings to understand the video by using cognitive skills. However, it's still a challenging task to develop an automatic system to fully understand the video content, although many approaches are proposed to detect high level events with semantic concept defined by human. This is because there is a large gap between low-level features and high-level semantics.

To bridge this gap, a mid-level representation framework is proposed in our previous work [1]. Under this framework, a mid-level representation is created between low-level features and high-level semantics. Some low-level features are used to classify every shot into semantic classes, such as "Player Close-up view", "Field View, "Audience view", "Player Medium View", etc. In [2], HMM network is used to divide the game into "Play" and "Break". In [3] audio keywords are created for the sports video, such as "Acclaim", "Commentator Speech", "Whistle" and "Silence" in soccer video.

Based on mid-level representation, some complex semantic events can be detected in the high-level layer. In [4], with the observation that in broadcast video there are additional information inserted into games such as multi-camera transactions and replays, some rules are extracted from the view type keyword sequence and events such as "Corner Kick" or "Goal" are detected using a decision tree built from these rules. In [1], some heuristic models are manually created to map mid-level keyword sequence to high-level semantic events.

However, there are some disadvantages with this kind of approach. Such rules or heuristic models might not be clear when the event has complex structure, and it is not easy to find and synchronize rules from different domain. Some systems take advantage of HMM based methods. In [5], a Sport-Game HMM (SG-HMM) framework is proposed. In the lower level of this framework, HMMs are used as quantizer to group low-level features. The sequences of these groups are then processed by higher level HMMs to detect event.

In this paper, we propose a framework to create audio/visual keywords by using low-level features from both visual and audio domains. A fusion scheme to synchronize and combine visual and audio keyword sequences is raised. Events are detected from the combined keyword vector sequence by using a HMM classifier which is able to automatically find temporal patterns, instead of defining heuristic rules manually.

This paper is organized as follows. Section 2 described the general structure of the event detection framework. Section 3 and 4 described the low-level to mid-level mapping and mid-level to high-level mapping in the framework, respectively. We apply our model to soccer game highlight extraction and the experiment result is given in section 5. With the result, conclusions are made and future works are raised in Section 6.

2. Our proposed model

The proposed model follows a 3-layer structure. In the low-level layer, video is divided into visual and audio streams. Features are then extracted and sent to the mid-

level layer. In the mid-level layer, Support Vector Machine (SVM) classifiers are used to group these features into predefined groups and each group is labeled with a keyword. Such keyword sequences are combined to form a keyword vector stream which is then processed in the high-level layer to detect semantic event. Figure 1 gives a diagram of this model framework.

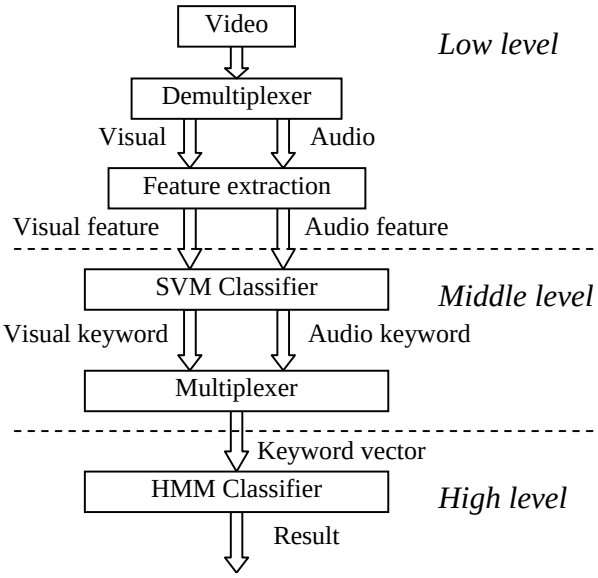

Figure 1: Diagram of our proposed framework

3. Low-level to mid-level mapping

3.1. Visual keyword creation

First, the motion intensity feature is used and a threshold is set empirically to divide the video stream into static and active parts. The static parts are represented by key frames, and the active parts are further segmented into sub-parts according to their color character. Then motion features (means and standard deviations of magnitudes and angles of motion vectors at block level and distribution of motion directions) are used to classify these sub-parts using SVM. As the original SVM is a two-class classifier, for multi-classes case (as in our model), SVM is used in "one-against-all" mode. Each sub-part is then labeled with a visual keyword. Detailed algorithm is described in our previous work [6].

For the events in a soccer game mentioned in our later experiment, the following visual keywords are defined: Far Middle-Field View, Far Goal-Field View, Close-Up View, Audience View, and Replay. These visual keywords have had some semantic perceptions, which facilitate the high-level semantic event detection.

The motivation that we classify the video in this manner is that, for broadcast video, certain view type changing pattern appears when some events happen. For example, "Shot" will always companies with a change from other view types (most probability from Far Middle-Field) to "Close-Up View", "Far Goal-Field View" or "Replay". And different event exhibits different changing pattern. Here as in the mid-level, enough keywords should be defined to explicitly describe different changing, thus enable the event detection in high-level of our framework.

Post process is applied to the visual key word created. Since the video frame rate is normally 25 to 30 fps, within a short time the view type should be equal. Sudden change in the keyword sequence is considered as error and post-processing is applied to eliminate this by using majority voting in a sliding window.

3.2. Audio keyword creation

In a broadcast sports game video like basketball or soccer, there are sounds like acclaim, commentator speech, whistle, silence, etc. These sounds are mainly made up of speech, environment noise and tool sound. We use the following features in our framework: Mel Frequency Cepstral Coefficients (MFCC), Liner Prediction Coefficient (LPC), LPC Cepstral (LPCC), Short Time Energy (STE), Spectral Power (SP), and Zero Crossing Rate (ZCR). However, for certain set of classes needed, only a few of these features are useful. An automatic audio feature selection algorithm described in our previous work [3] is used for feature selection. Since some of these features are multi-dimensions, the automatic algorithm will do both feature dimension selection and feature combination selection. For example, we defined four audio key words for the soccer game mentioned in our later experiment: "Acclaim", "Commentator Speech", "Whistle", and" Silence". To classify these four classes, LPC and MFCC features are found to be useful.

After the necessary low-level audio features are decided, the whole sound track from the game are framed using a sliding window of 20ms with no overlapping. For each of the audio frames, the decided features are extracted and another SVM classifier is used to give an audio keyword to the audio frame.

Like the visual keyword case, sudden change in audio keyword sequences is also considered as an error. Post process of audio keywords sequence is applied to remove the error and another sliding window of 5 frames with 1 frame overlap is used.

4. Mid-level to high-level mapping

Once the mid-level keywords are created, the high-level semantic event can be detected from such keyword sequences.

In event detection from keyword sequences, the heuristic approach needs to define rules for each event.. For example, if one rule is:

"Far mid-field view" to "Far goal-field view" to "Player close-up view" = "Shot"

Then in detection highlight, if this rule is matched, the sequence is thought to contain the event of "Shot".

The heuristic approach is not a statistical approach so it is sensitive to mid-level keyword creation error. Besides this, when a certain event has complex structure or has different view type change pattern, the rules are not easy to define. The HMM based classifier does not have these disadvantages due to two factors. Firstly, it is a statistical approach so it is more robust. Secondly, the temporal pattern can be automatically found by training.

4.1. HMM prototype definition

To detect the events in the experiment mentioned in Section 5, we define a generic HMM model (Prototype) for these events. For simplicity, we limit this model to a left-to-right HMM (later state cannot transit to earlier state). This limitation does not decrease the detection accuracy because events in sports video are from real world, time passes only from the early to the late.

Figure 2 illustrates the structure of the HMM prototype used in our framework.

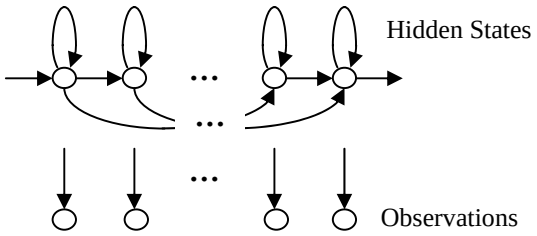

Figure 2: Left-to right prototype HMM structure

Figure 3 shows the structure of some sample visual keyword sequence used in training for soccer event detection. Each different view type keyword is an observation. The length of each view type is not the same, but this is not a problem for HMM which can divide the sequence using Viterbi Decoding. However, in the prototype enough states should be included so that it can explicitly model every event. We noticed that these sample sequences contain 3 to 6 different view types. If too few states are used, the detection accuracy will significantly drop as observed in our experiment. In our system we included 4 emitting states.

4.2. Fusion of keywords

The keywords are represented by index values of the codebook, so visual and audio keywords sequence is a one dimension stream each. Normally audio keywords are created based on a smaller sliding window size compared with visual frame rate, we synchronize audio keywords to visual keywords and a two-dimension keywords vector stream is then obtained and sent to HMM classifier for training and testing. The flow chart is shown in Figure 4.

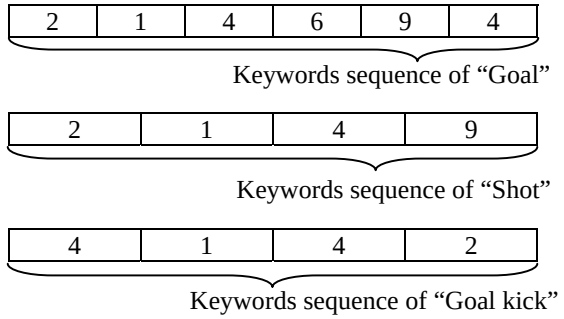

Figure 3: Visual keywords stream example.
1: Far goal-field view; 2: Far mid-field view;
4: Player close-up view; 6: Audience view;
9: Replay

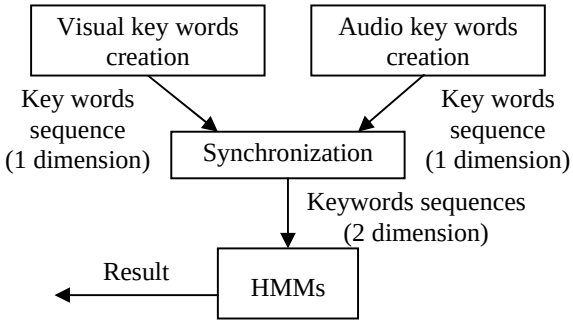

Figure 4: Flow chart of synchronization and fusion.

4.3. Unbiased data

Large amount of training sets should be obtained in training the HMM model for each event. But in some games, for example in soccer game, while the event of "Shot" or "Goal kick" happens quite often, the event of "Goal" happens rarely. The biased amount of training data will cause inaccuracy. In our experiment we noticed that if the HMM of "Goal" is not fully trained, this type of event will be recognized as "Shot". One way to solve this problem is to analyze more soccer games and collect enough training samples. An alternative way is to using the same training samples for more than one time.

5. Experiment

We tested our framework on broadcast soccer video. Soccer game has loose structure so detection of soccer events is more challenging. Totally 3 hours of FIFA

world cup 2002 game video is used in this experiment, We tried to detect the following events: “Goal (G)”, “Corner kick (C)”, “Shot (S)” and “Goal kick (GK)”.

5.1. Visual and audio keywords creation

In the visual domain, color, motion and texture are used to create the following keywords: “Far Middle-Field View (MF)”,”Far Goal-Field View (GF)”, “Close-Up View (C)”, “Audience View (A)”, and “Replay (R)”.

The visual key creation accuracy is shown in Table 1.

Class	*MF*	*GF*	*C*	*A*	*R*
Accuracy	*92%*	*94%*	*92%*	*95%*	*97%*

Table 1: Visual classification accuracy

In the audio domain, 30 minutes game audio is used in training, and the automatic feature selection algorithm decides the LPC and MFCC features to be used for the four audio classes: “Acclaim (A)”, “Whistle (W)”, “Commentator Speech (CS)” and” Silence (S)”.

The audio key creation accuracy is shown in Table 2.

Class	*A*	*W*	*CS*	*S*
Accuracy	*98.0%*	*97.3%*	*92.6%*	*91.1%*

Table 2: Audio classification accuracy

5.1. High level event detection result

After the keywords are created, the generated 2 dimension keyword vector sequences are sent to HMM classifier. The total experiment data are divided into 50%/50% for training/testing. We tried to detect the following events: “Goal (G)”, “Corner kick (C)”, “Shot (S)” and “Goal kick (GK)”.The final result is listed in Table 3 below:

Event	Detect	Miss	False alarm	Recall	Precision
G	5	0	4	100%	56%
S	32	7	1	82%	97%
C	18	2	1	90%	95%
GK	21	2	5	91%	81%

Table 3: Result of soccer events detection

Factors that affect the accuracy are studied by taking a look at the original video for some mis-classified cases. One factor is the accuracy of mid-level keyword creation. For example, we notice a case that, in a corner kick, player is kicking from the upper part of the screen and the camera is focusing on the goal field, which makes the whole corner kick process a single “Far Goal-Field View” throughout the corner kick. Another factor might affect the accuracy of event detection is that the insufficient amount of training data. The method mentioned in section 4.3 can only in some extent make up for this. The third factor is the prototype HMM definition. Result in Table 3 is based on a 6 states HMM prototype. We make a comparison for this factor as shown in Table 4.

Number of states	*5*	*6*	*7*
Over all accuracy	*75.0%*	*86.4%*	*85.2%*

Table 4: Accuracy with different states in HMM

6. Conclusions and future work

The experimental results have proved the using of a 3-layer event detection model can help to bridge the gap between low-level features and high-level semantic events. And the using of HMM can automatically find the temporal pattern of keyword streams instead of rules or heuristic models [1]. This gave the whole framework more flexibility. Using of SVM in mid-level had extended the limitation in HMM which, in modeling observation probabilities, only support Gaussian Mixture Model.

In the future, the accuracy of mid-level keyword creation can be improved to increase the overall event detection accuracy. Besides this, more visual and audio keywords can be defined to detect more events. Also keywords from different domains can be used so that events from other games, or even other type of videos, can be detected.

7. References

[1] L-Y Duan, M Xu, T-S Chua, Q Tian, C-S Xu, “A Mid-level Representation Frame-work for Semantic Sports Video Analysis”, In Proc. of ACM Multimedia' 03, ISBN:1-58113-722-2, Pages: 33-44 . Nov. 2003

[2] Xie, L., Chang, S-F., Divakaran, A., Sun, H., "Structure Analysis of Soccer Video with Hidden Markov Models", IEEE ICASSP, ISSN: 1520-6149, Vol. 4, pp. 4096-4099, May 2002

[3] J-J Wang, C-S Xu and E-S Chng, “A Generic Model for Automatic Audio Feature Selection in Sports Audio Classification”, submitted to ICASSP 2004.

[4] Y-L Kang, J-H Lim, Q Tian, M-S. Kankanhallis, “Soccer Video Event Detection with visual Keywords”, ICICS-PCM 2003, 15-18 December 2003, Singapore

[5] G Xu, Y-F Ma, H-J Zhang, S-Q Yang, “A HMM Based Semantic Analysis Framework for Sport Game Event Detection”, In Proc. of IEEE ICIP' 03, Vol. 1, Page(s): 25- 28

[6] H-P Sun, J-H Lim, Q Tian, M-S. Kankanhalli, “Semantic Labeling of Soccer Video”, ICICS-PCM 2003, 15-18 December 2003, Singapore