IDENTIFY SPORTS VIDEO SHOTS WITH "HAPPY" OR "SAD" EMOTIONS

Jinjun Wang[2,1], *Engsiong Chng*[2], *Changsheng Xu*[1], *Hanqing Lu*[3], *Xiaofeng Tong*[3]

[1] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
{stuwj2, xucs}@i2r.a-star.edu.sg
[2] CeMNet, SCE, Nanyang Technological University, Singapore 639798
aseschng@ntu.edu.sg
[3] NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China 100080
{hqlu, xftong}@nlpr.ia.ac.cn

ABSTRACT

Semantic video content extraction and selection are critical steps in sports video analysis and editing. The identification of video segments can be from various semantic perspectives, e.g. certain event, player or emotional state. In this paper, we examined the possibility of automatically identifying shots with "happy" or "sad" emotion from broadcast sports video. Our proposed model first performs the sports highlight extraction to obtain candidate shots that possibly contain emotion information and then classifies these shots into either "happy" or "sad" emotion groups using Hidden Markov Model based method. The final experimental results are satisfactory.

1. INTRODUCTION

Digital video editing and composition are common in film and broadcast post-production. The latest development in computer science has extended these technologies from professionals to home users. However, video editing remains a time-consuming and labor-intensive task because firstly, the interesting segment occurs infrequently in lengthy video documents, and secondly, good post-editing comes with skill, experience and artistic talent. For these reasons, the Automatic Video Editing problem attracts increasing research attention in recent years [1, 2, 3]. Intelligent video editing tools like muvee [4] are also available for both professional and amateur users. However, most current video analysis techniques rely on low-level video/audio features extraction [2, 3] while video editing requires high-level semantic understanding of the video content. The inability to efficiently map the low-level features to high-level semantics limits the performance of existing video editing systems. For example, the muvee software can generate music video that has good video/music boundary matching but the content of the video are not always interesting to the user.

In our previous work [5] we introduced an video editing system which automatically identifies interesting scenes of soccer event/player(s)/team from soccer video and match the video with music to produce music sports video (MSV). However, our system sometimes selects shots with "sad" emotion for "jolly" music portions, and thus the generated MSV is uncomfortable to the audience. This is due to the difficulty to evaluate the high-level emotional states of the shot from low-level video features. In fact the emotion analysis can provide very important information [6] for video editing tasks. This motivates us to examine the possibility to identify the shots emotions from broadcast sports video. Particularly, for the video editing purpose, we focus on recognizing shots with "happy" or "sad" emotions (Fig.1).

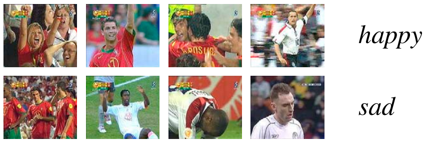

Fig. 1. Examples of emotional shots

In this paper we proposed a multi-level framework to identify shots with "happy" or "sad" emotion from the broadcast sports video. We use soccer video as an example because it is not only a globally popular game, it also presents many difficult challenges for video analysis due to its dynamic structure. The rest of the paper is organized as follows: Section 2 described the framework of our system, section 3 and section 4 explain the "Emotional shot candidate selection" and the "Emotion classification" modules in the framework, experimental results are listed in section 5 and section 6 discusses conclusions and future work.

2. OUR PROPOSED FRAMEWORK

Human emotions are reflected through facial expression, voice, hand and body gesture [6], and current researchers mainly

1-4244-0367-7/06/$20.00 ©2006 IEEE

perform emotion recognition by facial expression analysis [6] or speech analysis or a combination of the two. However, for broadcast soccer video the facial expression, voice and hand gesture information are not available because they are either not captured or because their qualities are too poor to be analyzed. This has made the emotional shot identification a challenging problem.

Despite the absence of some important information for emotion analysis, there is other domain-specific information available from soccer video, e.g. the broadcast soccer video production rules, the cinematic visual patterns, etc. To make use of these available cues to achieve satisfactory emotional shots identification result, a multi-level framework is proposed as shown in Fig.2. The highlight detection block extracts highlight scenes from the soccer video and the following processing blocks segment the highlight into shots and identify the "happy"/"sad" state in each shots. The following sections discuss the detailed procedures of the system.

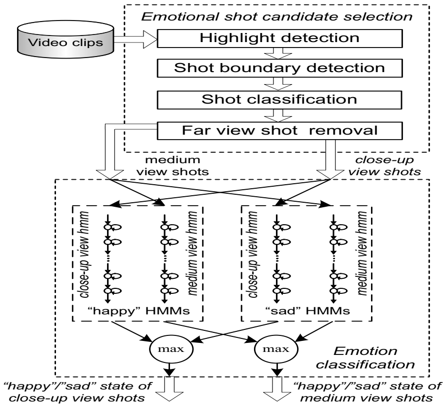

Fig. 2. System flow chart

3. EMOTIONAL SHOT CANDIDATE SELECTION

3.1. Soccer highlight detection

The broadcast soccer video contains consecutive shots, and these shot can be categorized into three types: far view, medium view and close-up view shot (Some researchers use an additional view class called "replay" view for soccer video. In fact, the "replay" view shots display previous game actions which are collections of far/medium/close-up view shots, hence the far/medium/close-up view classification is enough to cover all the view types for our research). From empirical observations of our soccer database, normally the emotion information can be identified during soccer highlight which consists of a series of events such as foul, injury, just-missing goal and goal-scoring. In these soccer events, one team would appear happy while the other team sad. Though there are other soccer events that do not possess emotion information, e.g. corner kick, goal kick, etc, usually these events would not be included in soccer highlight. For simplicity we use the term "event" to represent the soccer events included in soccer highlight in the following discussion. We also notice that the emotion information is mainly present from the medium/close-up view shots whereas the far view shots seldom tell any emotion information because the objects in this view are too small for analysis.

Based on these observations, we apply soccer highlight detection to filter out the shots that contain no emotion information. Sports event detection has been widely studied [7]. In this paper, we apply a robust highlights detection model from our previous work [8]. The key idea is to first extract dominant speech portions in the noisy commentary that corresponds to excited vocal, and then use the obtained dominant speech segments as the starting points to detect the quick succession of changing camera views which broadcasters typically use to showcase an interesting event.

3.2. Medium/close-up view shots extraction

In addition to highlight detection, the medium/close-up view shot segment information is identified from the detected highlight. This is achieved using the shot boundary detection and shot classification method described in our early work [9]. This method sequentially performs "Shot boundary detection", "Frame classification" and "Weighted majority voting" to obtain the shot boundary and classification information. In order to identify all the potential shots for emotion classification, the shot boundary detection and shot classification are performed over both the detected event segments and the neighboring segments at both ends whose length is set to be 0.5 times of the corresponding event segment length.

3.3. Emotional shot candidates extraction

After the highlight and shot classification information are obtained, potential emotional shots, i.e. the close-up/medium view shots, can be extracted. As the highlight detection method mainly focuses on the audio information, the detected event segments do not necessarily align with the visual information such as shot boundary. To properly combine the highlight detection with the shot boundary and classification information, a backward and forward alignment search is used.

To perform the alignment search, the common practice in current soccer broadcasting is firstly studied. In most broadcast soccer games, soccer events are first captured by the main camera which produces the far view shots of the broadcasting feed. Then during the successive game break by the event, the director would launch several sub-camera views to maximize the converge of the event. These sub-camera views would often be close-up/medium views showing the coach, players or audience reactions to the event as shown in Fig.1. These are the shots used by our system for emotion classification. When the match is restarted, the main camera view is then applied. In another word, the potential shot which is useful for emotion classification is often sandwiched by two far view shots.

Based on such “far view, close-up/medium view(s), far-view” structure in the neighborhood of the event segment, the close-up/medium view shots are identified using the backward and forward alignment search as illustrated by Fig.3. Specifically, the backward search is performed from the start of the event segment (the “e” region in the audio track in Fig.3) to locate the starting “far view” shot (the left “f” region in the visual track in Fig.3), and a similar forward search is performed to locate the ending far view shot. The medium/close-up shots found in this event segment are then applied to next module for emotion classification.

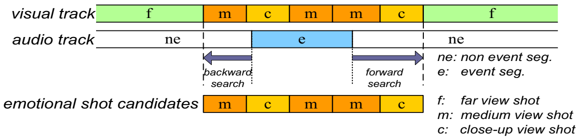

Fig. 3. Emotional shot candidates search

4. EMOTION CLASSIFICATION

Traditional techniques for emotion analysis would use facial expression, voice or hand gesture information. However, robust extraction of these features is extremely difficult in the captured close-up/medium shots from broadcast soccer video. Hence, we examine alternative features for emotion analysis, i.e., we use eight motion information, e.g. camera motion, motion entropy of the frame, motion magnitude of the motion vectors, etc, as features instead. Although these features do not directly reflect the emotion of persons or groups in the shot, our results show that there is a high correlation between these features and the two classes of emotion, “sad” and “happy”. The following sections describe the motion features and the classifier used in this paper.

4.1. Feature selection

The most salient difference between the “happy” and “sad” shots come from the behaviors of the object (i.e. coach, player and audience): A happy player would normally run or walk quickly which will result in high and diverse object motion. Alternatively, a sad player would normally stay still or walk slowly, resulting in small and uniform motion. These differences are reflected in the following features:

1) Camera motion: As the camera always follows the movement of the player, the camera motion provides a useful cue to represent the activity of the player. In our system, the following 6 features are extracted, specifically, “Average Motion Magnitude”, “Motion Entropy”, “Dominant Motion Direction”, “camera pan parameter”, “camera tilt parameter” and “camera zoom parameter”. These features are computed using the Motion Vector Field extracted from the compressed video.

2) Edge difference detection: The number of edge pixels and its difference between two consecutive frames can also provide a measure of the speed at which the camera view changes. These are the two edge motion features used.

Thus, an R^8 (camera motion R^6 + edge R^2) motion feature vector sequence is extracted from each candidate shot. In the next subsection, these sequences are classified to identify the emotion state to which the corresponding shots belong.

4.2. Emotion classification

To accurately classify the shot candidates, the motion features from individual frames as well as their temporal pattern from consecutive frames must be considered, and thus the HMM based classifier is suitable. The HMM based classifier has been shown to be robust and accurate for many problems [10], such as automatic speech recognition, image processing, communications, etc. Some recent work of HMM related to sports video analysis includes: Shinichi et al [11] proposed a HMM based system to classify different sports video, Xie et al [12] introduced a hierarchical HMM to analyze the soccer video structure and we [5] applied HMM for soccer event boundary detection.

For the emotion classification problem in this paper, our HMM prototype definition for both classes consists of 5 left-right HMM states with 3 gaussian mixture models in each state. For each class, two HMM models are created, one for medium view (*medium view HMM*) and the other for close-up view (*close-up view HMM*) as shown in Fig.2. The rationale for creating two separate HMMs for each class is to precisely model the patterns for the respective views in each class. During the testing process, the view shot close-up/medium state is used to select the appropriate HMM model for evaluation.

5. EXPERIMENTAL RESULTS

To examine the performance of our system, several experiments are conducted to evaluate each individual module in Fig.2. Firstly to measure the performance of the combined shot boundary detection and shot classification, three broadcast soccer videos (totally 2.5 hours) from key European leagues were used, and a shot classification accuracy of 94%

(89 incorrect in totally 1328 shots) was achieved. The performance of soccer highlight detection using only audio information was 81% using the method discussed in [8].

To test the performance of our emotion classification module, another 6 hour broadcast soccer video captured off live broadcast of key European leagues and Euro Cup 2004 was used. The obtained shots in the detected highlight segment are manually given a "happy" or "sad" label to serve as ground-truth to evaluate our emotion classification result. The accuracy is listed in Table 1.

Table 1. Emotion classification accuracy

Emotion class		Precision	Recall
Happy	Close-up view	85%	65%
	Medium view	90%	90%
Sad	Close-up view	68%	87%
	Medium view	82%	82%

The above results showed that the "sad" close-up view shots resulted in the worst precision. The analysis of the errors for these misidentifications showed that these "sad" shots contain frames with high motion in the close-up view and/or surrounding object motion, e.g. movement of unrelated person coming into the camera view.

To improve classification performance for the close-up view, we include a face motion detection module to improve performance. Our strategy is based on the observation that in a sad close-up view, despite the interference of surrounding object motion, the sad player's face can often be detected and tracked. Examples of "sad" faces captured in the close-up are shown in Fig.4.

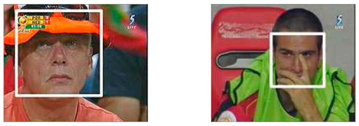

Fig. 4. Face detection for "sad" class

Hence we apply an additional module to re-score the close-up view HMM. Specifically, if a face is detected and has a motion activity score lower than a specified threshold, it is classified as a "sad" event, otherwise the HMM result is retained. This additional module improves our "sad" close-up view precision to 79%.

6. CONCLUSION AND FUTURE WORK

In this paper we describe the use of the "happy"/"sad" semantic for sports video analysis and the use of motion and edge information to identify shots with "happy"/"sad" emotional state from broadcast soccer video.

The ability to classify shots with emotion information greatly facilitates video content analysis. Our future work includes: 1) Investigating more features and analysis methods to improve the performance of the system besides the method discussed in section 5, 2) Increasing the HMM complexity to have more models per class. 3) Examining the possibility to extend the system to other sports domains such as tennis and basketball, and 4) Concatenating the emotional shot identification system in this paper to other modules to perform the automatic sports video editing task.

7. REFERENCES

[1] Xiansheng Hua, Lie Lu, and Hongjiang Zhang, "Ave: automated home video editing," *Proc. of ACM MultiMedia'03*, pp. 490–497, 2003.

[2] A. Girgensohn, et al, "A semi-automatic approach to home video editing," *Proc. of UIST'00, ACM Press*, pp. 81–89, 2000.

[3] Jonathan Foote, et al, "Creating music videos using automatic media analysis," *Proc. of ACM MultiMedia'02*, pp. 553–560, Dec. 2002.

[4] MuVee Technologies Pte. Ltd, "Muvee^{TM}," 2000.

[5] Jinjun Wang, et al, "Automatic generation of personalized music sports video," *Proc of ACM MultiMedia'05*, pp. 31–38, November 2005.

[6] Thomas S. Huang, "Emotion recognition using a cauchy naive bayes classifier," *Proc. of IEEE ICPR'02*, 2002.

[7] N Adami, et al, "An overview of multi-modal techniques for the characterization of sport programmes," *Proc. of SPIE-VCIP'03*, pp. 1296–1306, July, 2003.

[8] Kongwah Wan and Changsheng Xu, "Robust soccer highlight generation with a novel dominant-speech feature extractor," *Proc. of IEEE ICME'04*, 2004.

[9] Jinjun Wang, Engsiong Chng, and Changsheng Xu, "Soccer replay detection using scene transition structure analysis," *Proc. of IEEE ICASSP'05*, March 2005.

[10] Valery A. Petrushin, "Hidden markov models: Fundamentals and applications," *Online Symposium for Electronics Engineer*, 2000.

[11] S. Takagi, et al, "Sports video categorizing method using camera motion parameters," *in Proc. of ICME'2003*, July 2003.

[12] Lexing Xie, et al, "Structure analysis of soccer video with domain knowledge and hidden markov models," *Pattern Recognition Letters*, vol. 24, December 2003.