A Two-Stage Approach to Highlight Extraction in Sports Video by Using AdaBoost and Multi-modal

Shaojie Cai[1,2], Shuqiang Jiang[2], and Qingming Huang[1,2]

[1] Graduate University of Chinese Academy of Sciences, Beijing, China
[2] Key Lab of Intell.Info.Process., Inst.of Comput. Tech., Chinese Academy of Sciences, Beijing, China
{sjcai,sqjiang,qmhung}@jdl.ac.cn

Abstract. In this paper, we propose a novel two-stage approach for highlight extraction in sports video. In the first stage, a preliminary classification is performed to the audio stream to locate the position of the highlight candidates. We employ AdaBoost algorithm for feature selection and audio classification. In the second stage, we extract visual and temporal features of these highlight candidates and feed them into a linear weighted model for further highlight extraction. The final highlight segments are determined based on the output value of the model. The advantage of this method is its low computational complexity and relatively high accuracy. Experimental results on tennis video demonstrate effectiveness and efficiency of our proposed approach.

Keywords: Highlight extraction, AdaBoost, feature selection, audio classification, Linear Weighted Model.

1 Introduction

In recent years, content-based highlight extraction for sports video has been a hot topic and developed rapidly and extensively. Researches have studied the topic mainly based on visual or audio modalities [1, 2].

In this paper, we present a novel highlight extraction approach by using AdaBoost and multi-modal. The approach is motivated by two facts. Firstly, single-modality cannot fully represent feature interpretation of video data. Secondly, to alleviate computational cost, we employ AdaBoost to select the most critical features and perform audio classification; In addition, we apply two-stage approach to filter out the highlight-irrelevant segments at the first stage, hence greatly reduce computational cost at the following decision stage.

Our proposed system is shown in Fig.1, including two major parts: 1) preliminary audio classification; 2) highlight extraction based on visual-temporal fusion.

As mentioned above, our system has two main features: 1) utilize AdaBoost algorithm to select audio features and train classifiers for audio analysis; 2) focus on audio analysis and use visual and temporal information as auxiliary to improve the overall accuracy.

Y.-M.R. Huang et al. (Eds.): PCM 2008, LNCS 5353, pp. 867–870, 2008.

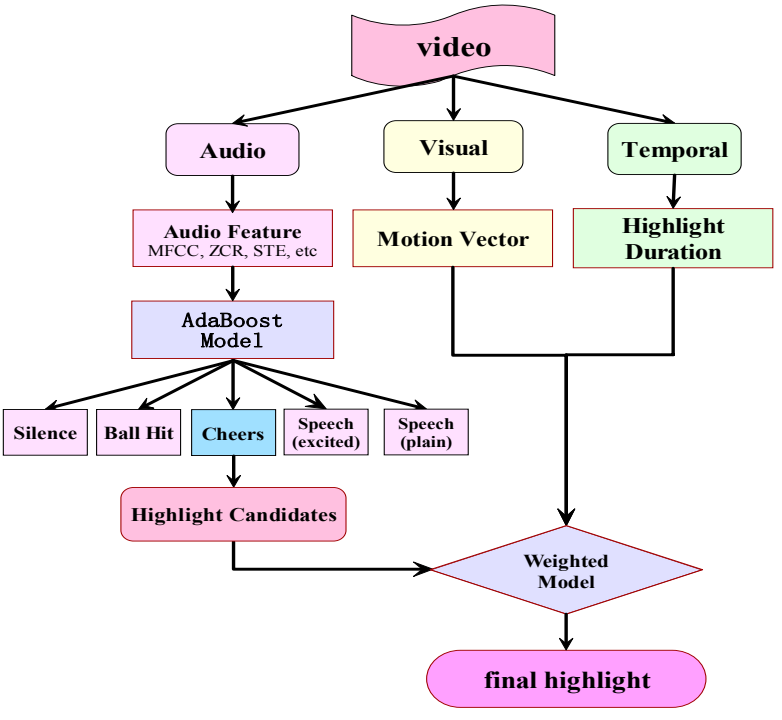

Fig. 1. System Framework

2 First Stage: AdaBoost-Based Audio Classification

In sports video, audio stream generally includes speech, music, audience sound and various environmental sounds. In this paper, we will classify five different sounds, which is ball-hit, cheers, silence, excited commentator speech and plain commentator speech. Generally speaking, audiences tend to express their excitement with louder voice during exciting moments, and duration of cheers is usually longer. Therefore, we focus on the cheers as the audio cue for further highlight extraction work.

We segment the audio at 1s per clip as the basic unit for feature extraction. Short-time energy (STE), Zero-crossing Rate (ZCR), Spectrum Flux, pitch, brightness, bandwidth, sub-band power, LPCC and MPCC are selected as low-level audio features. A total of 55 features are extracted from each clip.

We choose AdaBoost as feature selector and audio signal classifier. AdaBoost is an adaptive algorithm to boost a sequence of weak classifiers, in which the weights are updated dynamically according to the errors in the previous learning [4]. AdaBoost has arisen wide attention due to its excellent performance in many pattern classification problems such as image and video content analysis and annotation. However, AdaBoost have not been well explored in the audio domain of sports video analysis. We evaluate AdaBoost algorithm of Tieu and Viola's version [5] for audio classification. The algorithm made the weak learner work with a single feature at a time.

Therefore, the most critical features can be simultaneously selected in the learning process. Above AdaBoost algorithm is only for two-class classification. In this case, we use "one against all" strategy and decision tree as weak classifier to get the final decision.

3 Second Stage: Highlight Extraction

In order to extract highlight segments from the whole sports video effectively and rapidly, we take two stages to complete the task. Firstly, a preliminary classification is performed to the audio stream to locate the position of "highlight candidate", i.e., beginning and end frame of the segments. We locate the segment labeled "cheers" extracted from previous audio classification work and define them as "highlight candidates". Then, a linear weighted model is developed using visual and temporal features extracted from highlight candidates. In the visual dimension, we employ the average motion vector (*AMV*) in 1 second, defined as

$$AMV = \sum_{i=1}^{k} MV_i / k \tag{1}$$

where MV_i is the motion vector of the i th frame, k is the frame rate of video.

In the temporal dimension, we employ cheers duration (*CD*). Both the visual and temporal feature are normalized, defined as v_{amv} and T_{cd}. Then, we feed them into a linear weighted model with proper weights w_{mv} and w_{cd} that sum up to 1.0. Each highlight candidates' output value is compute as:

$$HD = w_{mv} \cdot V_{amv} + w_{cd} \cdot T_{cd} \tag{2}$$

We select the segments whose *HD* values are above threshold *conf* as final highlights.

4 Experimental Results

The experimental database comprises five tennis video clips extracted from five living broadcast programs of French Open 2005. The database is partitioned into 66%/34% training/testing set.

For audio classification, 7 effective features out of 55 features are selected to form the reduced feature vector by using AdaBoost. To test the performance of AdaBoost as a feature selector and classifier, we use single feature, total 55 dimensional features and selected features to train and test AdaBoost respectively. In addition, SVM classifier with total 55 dimensional features is also tested. Table 1 shows classification accuracy of cheers detection. We can see that AdaBoost performs better than SVM classifier with higher precision rate and recall rate. Precision rate slightly declines by using reduced feature set. However, considering considerable reduction of computational cost, the result is quite satisfactory.

For highlight extraction, we firstly apply a method to set up ground truth: four persons are invited to give each highlight candidate a score limited from 0 to 1. The final

score of each segment is the average of the four values. We select the segments whose scores are above 0.6 as ground truth.

We denote the number of final highlight as N, the number of Candidates as the C, and the number of highlights in the ground truth as G, then precision is defined as the ratio between N and C , i.e., $P = N/C$, recall is defined as the ratio between N and G, i.e., $R = N/G$. Experimental data and testing results are shown in Table 2. It can be seen that our scheme has achieved satisfactory results.

Table 1. Classification Accuracy for Detecting Cheers

	Precision	Recall
ZCR	46.2%	48.0%
MFCC	78.0%	80.2%
SVM+55	86.7%	88.0%
Ada+55	88.9%	96.0%
Ada+7	80.0%	96.0%

Table 2. Experimental Result for Highlight extraction

Video	Precision	Recall
No.1	69.6%	76.2%
No.2	71.9%	74.2%
No.3	74.5%	77.4%
No.4	70.5%	72.9%
No.5	70.8%	73.2%

5 Conclusions

A novel two-stage highlight extraction approach based on audio, visual and temporal modalities is proposed in this paper. By employing AdaBoost to select audio features and classify video clips to locate highlight candidates, and use motion and temporal information as auxiliary to obtain the final highlights, our method has achieved relatively high accuracy, while greatly reduce computational complexity compared to the existing method. We will further explore more effective features and statistical learning algorithms to boost the performance of the approach.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China under Grant 60773136 and 60702035, in part by National Hi-Tech Development Program (863 Program) of China under Grant 2006AA01Z117.

References

1. Rui, Y., Gupta, A., Acero, A.: Automatically extracting highlights for TV baseball programs. In: Eighth ACM International Conference on Multimedia, pp. 105–115 (2000)
2. Xie, L., Chang, S.F., Divakaran, A., Sun, H.: Structure analysis of soccer video with hidden Markov models. In: Proceedings of the international conference on acoustic, speech, and signal processing, vol. 4, pp. 4096–4099 (May 2004)
3. Hanjalic, A.: Generic approach to highlight detection in a sport video. In: Proceedings of the IEEE international conference on image processing, vol. 1, pp. 1–4 (September 2003)
4. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of online learning and an application to boosting. Journal of Computer and System Sciences 55(1), 119–139 (1997)
5. Tieu, K., Viola, P.: Boosting image retrieval. In: Proc. of Computer Vision and Pattern Recognition, vol. 1, pp. 228–235 (2000)