MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Generation of Sports Highlights Using Motion Activity in Combination with a Common Audio Feature Extraction Framework

Xiong, Z.; Radhakrishnan, R.

TR2003-118 September 2003

Abstract

In our past work we have used temporal patterns of motion activity to extract sports highlights. We have also used audio classification based approaches to develop a common audio-based platform for feature extraction that works across three different sports. In this paper, we combine the two aforementioned complementary approaches so as to get higher accuracy. We propose a framework for mining the semantic audio visual labels in order to detect interesting events. Our results show that the proposed techniques work well across our three sports of interest, soccer, golf and baseball.

GENERATION OF SPORTS HIGHLIGHTS USING MOTION ACTIVITY IN COMBINATION WITH A COMMON AUDIO FEATURE EXTRACTION FRAMEWORK

Ziyou Xiong, Regunathan Radhakrishnan, Ajay Divakaran

Mitsubishi Electric Research Laboratories, USA
Email: {zxiong, regu, ajayd}@merl.com

ABSTRACT

In our past work we have used temporal patterns of motion activity to extract sports highlights. We have also used audio classification based approaches to develop a common audio-based platform for feature extraction that works across three different sports. In this paper, we combine the two aforementioned complementary approaches so as to get higher accuracy. We propose a framework for mining the semantic audio-visual labels in order to detect "interesting" events. Our results show that the proposed techniques work well across our three sports of interest, soccer, golf and baseball.

1. INTRODUCTION

Past work on automatic extraction of highlights of sports events in general and soccer video in particular has relied on color feature extraction, audio feature extraction as well as camera motion extraction (see [1] [2] for example). Color and texture features have been employed to find the video segments in which the background is mostly green i.e. mostly consists of grass. Another approach has been to detect goalposts. Camera motion has also been shown to correlate with interesting events in soccer video. Audio features have been used to detect interesting events by looking for an increase in the volume or of the pitch of the commentator's voice or by even looking for the word "goal" in the commentary. Finally, combinations of the various features have been used to get refined results.

While the aforementioned techniques are simple, those that depend on the camera motion require accurate motion estimation for their success. Since our motivation is computational simplicity and easy incorporation into consumer system hardware, we focus on feature extraction in the compressed domain. In the compressed domain, however, since the motion vectors are noisy, such accuracy is difficult to achieve. In our previous work [3] we have shown that it is possible to rapidly generate highlights of various sports using gross motion descriptors such as the MPEG-7 motion activity descriptor. Such descriptors work well with compressed domain motion vectors, since they are coarse by definition. However, we found that with soccer video, the number of false positives was unacceptably high. In [4] we eliminated false positives by first detecting sudden surges in audio volume or peaks, and then only retaining the motion activity based highlights that correspond to an audio peak. While this procedure works reasonably well, it does not use a reliable audio feature. Thus, in [5] we developed an audio-classification based approach in which we explicitly identify applause/cheering segments, and use those to identify highlights. Thus, in this paper, we focus on using gross motion features such as the MPEG-7 motion activity descriptor in conjunction with audio classification to generate sports highlights, so as to maintain the computational simplicity while enhancing the accuracy. We also use our audio classification framework to set up a future investigation of fusion of audio and video cues for sports highlights extraction.

2. DESCRIPTORS OF MOTION ACTIVITY

A human watching a video or animation sequence perceives it as being a slow sequence such as a "news reader shot", or a fast paced sequence such as "goal scoring moment" or an action sequence etc. The MPEG-7 [3] motion activity descriptor captures this intuitive notion of 'intensity of action' or 'pace of action' in a video segment. The MPEG-7 intensity of motion activity descriptor is extracted by suitably quantizing the variance of the motion vector magnitude to one of 5 possible levels – very low, low, medium, high and very high. In our previous work, we have shown that the average motion vector magnitude can also be suitably quantized to the same kind of scale but is slightly inferior in fidelity to the ground truth. During our work, we found that the motion vector magnitude works well as an indicator of motion activity in the limited context of a soccer game and hence we choose to use it for computational simplicity.

3. AUDIO CLASSIFICATION FRAMEWORK

The system constraints of our target platform rule out having a completely distinct algorithm for each sport and

motivate us to investigate a common unified highlights framework for our three sports of interest, golf, soccer and baseball. Since audio lends itself better to extraction of content semantics, we start with audio classification. In [5] we describe an audio classification based framework illustrated in the left branch in Figure 1. We employ a general sound recognition framework based on Hidden Markov Models (HMM) using Mel Frequency Cepstral Coefficients (MFCC) to classify and recognize the following audio signals: applause, cheering, music, speech and speech with music. The former two are used for highlights extraction due to their strong correlation with highlights and the latter three are used to filter out the uninteresting segments. While the above technique is promising, we find that it still has room for improvement as can be seen in Table 1. First, the classification accuracy needs to be improved. Second, using applause duration alone is probably simplistic. Its chief strength is that it uses the same technique for three different sports.

4. PROPOSED TECHNIQUE

Since we do not expect a high gain from increased classification accuracy alone, we are motivated to combine visual cues with the audio classification with the hope that we may get a bigger gain in highlight extraction efficacy. In [3] we find that golf has a clear pattern of sharply rising motion activity corresponding to shots. We illustrate motion activity patterns corresponding to soccer and golf in Figure 2. We investigate combination of audio classification with the motion activity pattern matching. We illustrate our general framework in Figure 1. Note that the audio classification and the video feature extraction both produce candidates for sports highlights. We then propose to use probabilistic fusion to choose the right candidates. Note also that the video feature extraction goes well beyond the motion activity patterns that we described earlier.

For audio classification, unlike what we did in [5], in this paper we use Gaussian Mixture Models(GMM) instead of HMM to further reduce computation complexity. Also we introduce a "none-of-the-above" class to avoid the problem of inevitable errors using a closed set of classes when some signals do not belong to the closed set. Silence segments are declared if the energy of the segment is no more than 10% of the average energy of all the segments in the whole game. We then classify every second of non-silence audio into the following 5 classes: audience reaction sound(including applause, cheering etc), music, speech, speech with music and none-of-the-above(including ball-hits in gold, ball-bat impact in baseball and whistle in soccer etc).

Once we have obtained semantic labels for both audio and motion from sports video, the problem of highlight extraction can be posed as a multimedia mining problem across these labels for patterns. Figure 3 illustrates the proposed framework for fusion of audio and motion activity labels. The basic assumption in this approach is that interesting events are “rare” and have different audio and motion characteristics in time, when compared to the “usual” characteristics. In order to quantify what is considered as “usual”, we compute the joint two dimensional histogram of motion activity and audio labels, within a time window of length W_L. Then, for every smaller time window W_S within W_L , we compute the same joint distribution. We compare the local statistic (computed within W_S) with the global statistic (computed within W_L) using an information theoretic measure called relative entropy. We compute the average of all relative entropy values within W_L. One would expect a large relative entropy value for a W_S with a different distribution compared to what is “usual” within W_L. By moving W_L one W_S step at a time, we compute the relative entropy values throughout the sports video. Then, interesting rare events are times when there is a local maximum in the averaged relative entropy values.

In the following section, we present the audio classification results as well as the results of the proposed fusion approach.

5. EXPERIMENTAL RESULTS

5.1 Results of GMM+MFCC on a Small Dataset

To train the GMM using MFCC features, we’ve collected 1500 1-second long audio clips from TV broadcasting of golf, baseball and soccer games. Each of them is hand-labeled into one of the 5 classes: audience reaction sound, music, speech, speech with music and none-of-the-above. Each class has 300 clips. The audio signals are all mono-channel, 16 bit per sample with a sampling rate of 16kHz. The database is partitioned into a 70%/30% training/testing set. Each GMM uses 32 mixtures and for every 30ms of audio a feature vector with 24 elements is extracted(12 MFCC and 12 delta MFCC). The learned GMM are used to test on both the 70% training data and 30% of the test data. The recognition matrices are in Table 2 and 3 respectively. From these two tables, we can see the classifiers do very well on the training data while slightly worse on the test data. The none-of-the-above class is hard to learn to generalize well on unseen data.

5.2 Results on Classifying 3 Games

We use the trained GMM model to classify a 1-hour long baseball game, a 2-hour long golf game and a 2-hour long soccer game. We establish their ground truth by labeling each 1-second long audio segment. We compare the classification results and the ground truth and the recognition matrices are shown in Table 4~6.

The classification results drop even more on the almost unconstrained game soundtracks than on the 30% test data in the previous sub-section. The classification on the soccer game is worse than the other two games. The fourth row of Table 6 suggests that many speech segments(including noisy speech) are mis-classified as audience reaction sound such as cheering or loud noise. We will address this problem in the future.

5.3 Results on Combining Audio with Video Features

From the input sports audio track, every second of audio was classified into one of the trained sound classes. Motion activity values for P frames in every second of video were averaged and quantized using MPEG-7 thresholds. In order to extract highlights in the proposed framework, we chose a value of 3 minutes for W_L and 1 minute for W_S. The figure 4 shows the plot of averaged relative entropy values for a soccer video. The local maxima are times declared as "interesting" by our algorithm. We found that both the "goals" were detected among other "rare" events. The time stamps for which audio labels are "speech with music" are commercials and can simply be eliminated from highlights.

6. CONCLUSION AND FUTURE WORK

Our proposed techniques have the advantage of simplicity and fair accuracy. In ongoing work, we are examining more sophisticated methods for audio-visual feature fusion, better ways of representing highlights across the three different sports besides using the capture of goals as a measure of the accuracy. We are currently working on a framework to assess the highlights in terms of user satisfaction, so as to get a more complete assessment.

7. REFERENCES

[1] S. A. Dagtas and M. Abdel-Mottaleb, "Extraction of Soccer Highlights using Multimedia Features," *MMSP*, 2001

[2] C. Toklu, S-P. Liou and M.Das, "Video Abstract: A Hybrid Approach to Generate Semantically Meaningful Video Summaries, *IEEE ICME*, New York, 2000.

[3] K.A. Peker, R. Cabasson and A. Divakaran, "Rapid Generation of Sports Highlights using the MPEG-7 Motion Activity Descriptor," SPIE Conference on Storage and Retrieval from Media Databases, San Jose, CA, USA, January 2002.

[4] R. Cabasson and A. Divakaran, "Automatic Extraction of Soccer Video Highlights using a combination of motion and audio features," *SPIE Conference on Storage and Retrieval for Media Databases*, Santa Clara, CA, USA, January 2003.

[5] Z. Xiong, R. Radhakrishnan, A. Divakaran and T. Huang, "Audio Events Extraction based Highlights Extraction from Baseball, Golf and Soccer Games in a Unified Framework," to appear in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, April 2003.

Figure 1: Highlights Extraction Framework: The Audio part is described in [5]. We have partially realized the video and the probabilistic fusion so and plan to present further results at the conference. The learning machine is a HMM in this case.

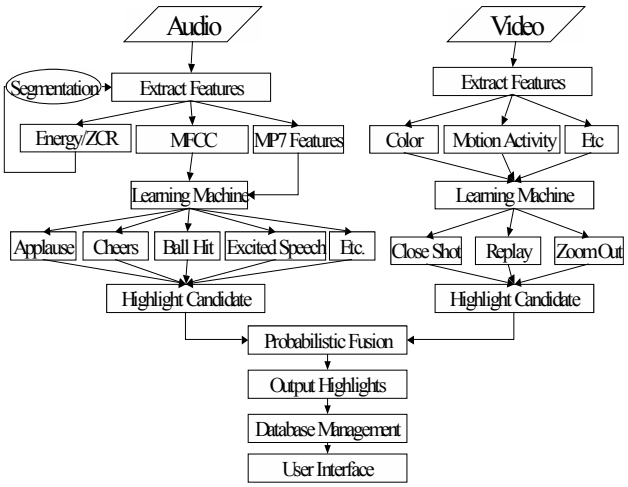


Figure 2: Activity curves for the golf(above) and soccer game(below)

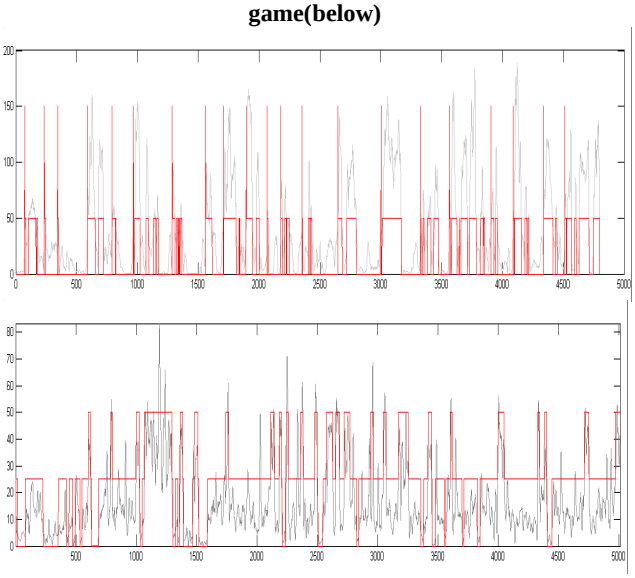


Figure 3: Framework for fusion of motion activity and audio labels for highlight extraction

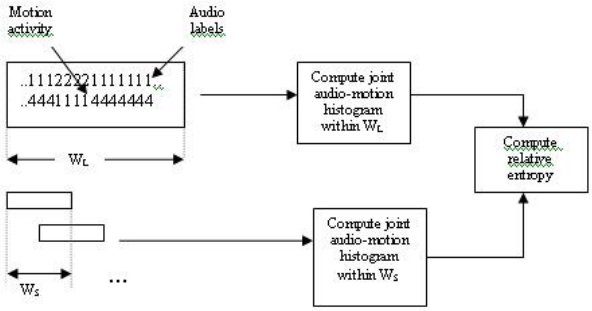
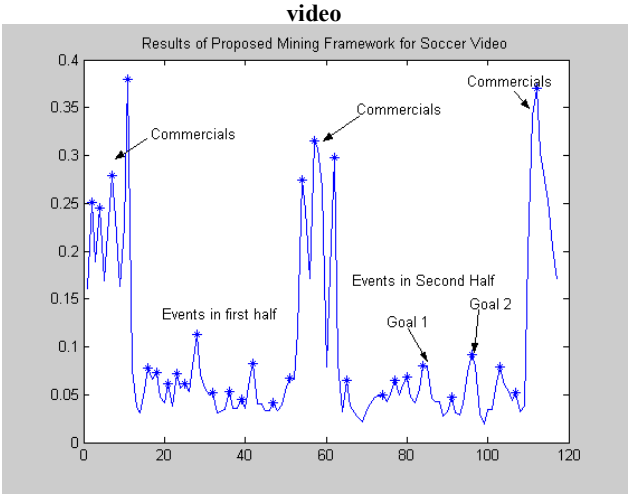


Figure 4: Results of the mining framework for a soccer video



	A	B	C	D	E	F	G
[1]	58	47	35	60.3	74.5	151	23.1
[2]	42	94	24	57.1	25.5	512	4.7
[3]	82	290	72	87.8	24.8	1392	5.2
[4]	54	145	22	40.7	15.1	1393	1.6

Table 1: This is the result reported in [5] on [1] golf game 1; [2] golf game 2; [3] baseball game; [4] soccer game. [A]Number of Applause and Cheering Portions (NACP) in Ground Truth; [B] NACP WITH Post-procesing; [C] Number of TRUE ACP by Classifiers; [D] Precision C/A; [E] Recall C/B WITH Post-processing; [F] NACP WITHOUT Post-processing; [G] Recall C/F WITHOUT Post-Processing

	[1]	[2]	[3]	[4]	[5]
[1]	0.995	0	0.005	0	0
[2]	0.005	0.995	0	0	0
[3]	0	0.006	0.962	0.019	0.103
[4]	0	0.005	0	0.995	0
[5]	0	0	0	0.005	0.995
overall	0.9886				

Table 2: Classification Results of the GMM on the training data(70% of 300 1-sec clips for each class). [1] audience reaction sound; [2] music; [3] none-of-the-above; [4] speech [5] speech with music.

	[1]	[2]	[3]	[4]	[5]
[1]	0.9978	0	0.011	0	0.011
[2]	0.033	0.711	0.022	0.011	0.222
[3]	0.06	0.09	0.657	0.06	0.134
[4]	0.011	0.011	0.011	0.9	0.067
[5]	0	0.033	0	0.011	0.956
overall	0.8402				

Table 3: Classification Results of the GMM on the 30% test data(30% of 300 1-sec clips for each class).

	[1]	[2]	[3]	[4]	[5]
[1]	0.5209	0.0133	0.1162	0.2803	0.0418
[2]	0	0.8387	0.0081	0.0161	0.1048
[3]	0.4167	0.0313	0.2188	0.2708	0.0625
[4]	0.0872	0.0064	0.0035	0.8239	0.0778
[5]	0.0029	0.0526	0.0015	0.2164	0.7237
overall	0.707				

Table 4: Classification Results of the GMM on the 1 hour long baseball game. Notice that the sum of each row is not necessarily 1 because false classification into silence is not put into the table. But the overall accuracy as in the last row takes this into account.

	[1]	[2]	[3]	[4]	[5]
[1]	0.1385	0.0229	0.0459	0.0798	0.0771
[2]	0.0150	0.7191	0.0075	0.0599	0.1873
[3]	0.0476	0.0952	0.2381	0.2619	0.1548
[4]	0.0031	0.0059	0.0098	0.8198	0.1320
[5]	0.0040	0.0714	0.001	0.1166	0.8070
Overall	0.6630				

Table 5: Classification Results of the GMM on the 2-hour long golf game.

	[1]	[2]	[3]	[4]	[5]
[1]	0.9537	0.0057	0.0249	0.0093	0.005
[2]	0.1312	0.7421	0.0204	0.009	0.095
[3]	0.4655	0.1336	0.056	0.1164	0.2241
[4]	0.4395	0.0064	0.0014	0.1934	0.3590
[5]	0.0464	0.0639	0.0015	0.1292	0.7591
Overall	0.4273				

Table 6: Classification Results of the GMM on the 2-hour long soccer game.