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

Generation of Sports Highlights Using a Combination of Supervised & Unsupervised Learning in Audio Domain

Radhakrishan, R.; Xiong, Z.; Divakaran, A.; Ishikawa, Y.

TR2003-144 February 2004

Abstract

In our past work we have used supervised audio classification to develop a common audio-based platform for highlight extraction that works across three different sports. We then use a heuristic to post-process the classification results to identify interesting events and also to adjust the summary length. In this paper, we propose a combination of unsupervised and supervised learning approaches to replace the heuristic. The proposed unsupervised framework mines the semantic audio-visual labels so as to detect "interesting" events. We then use a Hidden Markov Model based approach to control the length of the summary. Our experimental results show that the proposed techniques are promising.

Generation of Sports Highlights Using a Combination of Supervised & Unsupervised Learning in Audio Domain

Regunathan Radhakrishan[1], Ziyou Xiong[1], Ajay Divakaran[1], Yasushi Ishikawa[2]
[1]Mitsubishi Electric Research Labs, Cambridge, MA, USA.
{regu, zxiong, ajayd}@merl.com
[2]Mitsubishi Electric – Information Technology Research Center, Ofuna, Kamakura, Japan.
yasushi@isl.melco.jp

Abstract

In our past work we have used supervised audio classification to develop a common audio-based platform for highlight extraction that works across three different sports. We then use a heuristic to post-process the classification results to identify interesting events and also to adjust the summary length. In this paper, we propose a combination of unsupervised and supervised learning approaches to replace the heuristic. The proposed unsupervised framework mines the semantic audio-visual labels so as to detect "interesting" events. We then use a Hidden Markov Model based approach to control the length of the summary. Our experimental results show that the proposed techniques are promising.

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. In [9], Ekin et al propose an approach that detects dominant color regions and some other low-level video features to detect goals, referee and penalty boxes. In [8], Pan et al propose detection of slow motion segments as a solution to detect "interesting" events in sports 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. In [10], Xu et al propose a audio classification framework based on SVMs (Support Vector Machines) to detect whistles, commentators excited speech etc. The proposed method relies on rules based on these low-level audio events to extract highlights. While most of the above approaches use supervised learning for event detection, Xie et al propose a completely unsupervised approach to discover play and break segments in soccer video in [7]. Finally, combinations of the various features have been used to get refined results.

While most of the aforementioned techniques rely on video domain techniques, they are computationally intensive than those that rely on audio domain techniques. 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 [3] we also 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, we developed an audio-classification based approach in which we explicitly identify applause/cheering segments, and use those to identify highlights. In [4] we post-process the classification results to identify "interesting" events. Specifically, we use length of contiguous applause/cheering audio segments to rank the detected highlight events thereby enabling modulation of summary length.

In this paper, we propose a combination of unsupervised and supervised learning in place of the post-processing step in [4]. The unsupervised approach relies on the departure from stationarity of audio labels histogram to detect candidate "interesting events". We then further test the validity of these events using a pre-trained highlight Hidden Markov Model.

In section 2, we describe our motivation for the chosen audio classification based framework. In Section 3, we give details on the proposed techniques. In Section 4, we present the experimental results and conclude in Section 5.

2. Motivation for Audio Processing

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 [3]

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. This kind of processing based on low-bandwidth audio signal lends itself towards a platform for the analysis of different sports video followed by more sophisticated game-specific post-processing.

3. Proposed Techniques

3.1 Audio Classification Framework

In this paper, we follow the approach in [4] using Gaussian Mixture Models(GMM) to model classes of sound. 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 7 classes: applause, ballhits, female, male, music, speech-music and noise (audience noise including cheering).

3.2 Unsupervised Label Mining Framework

Once we have obtained semantic labels for audio from the GMM, we pose the problem of "interesting" event detection as a multimedia mining problem across these labels for patterns. Figure 1 illustrates the proposed framework for mining audio labels. The basic assumption in this approach is that interesting events are "rare" and have different audio characteristics in time, when compared to the "usual" characteristics in a given context. A context is defined by a time window of length W_L. In order to quantify what is considered as "usual" in a context, we compute the distribution of labels within it. Then, for every smaller time window W_S, we compute the same distribution. We compare the local statistic (computed within W_S) with the global statistic (computed within W_L) using either an information theoretic measure called relative entropy or a histogram distance metric proposed in [6]. One would expect a large distance value for a W_S with a different distribution compared to what is "usual" within W_L. By moving W_L one W_S at a time, we compute (W_L / W_S) distance values and find the maximum of this set of values. We associate this maximum value, M_{ws}, with the small window W_S. Then, rare events are times when there is a peak in the curve of all M_{ws}.

For instance, in a golf game the onset of audience applause would cause the local distribution of semantic audio labels to peak around applause whereas the global distribution in the current context, would peak around speech. Therefore, a large difference between these two distributions would indicate a deviation from stationarity, thereby signaling the occurrence of something "unusual" in that context.

The peak detection in the curve of all Mw_s, is performed adaptively using a sliding window similar to shot detection in [5]. When the first local maximum value is at the center of the window, we compare if its value is greater than the second local maximum value by a factor P_{th}. This helps eliminate peaks that are a result of random fluctuations.

3.3 Highlights Validation by Trained HMM

Points of unusual events detected by the above algorithm only signal a change in characteristics in terms of the semantic audio labels. For instance, in a golf game, the onset of commercials would also be signaled as an unusual event. Therefore, in order to capture only semantic events of interest, we use a trained HMM of highlights to validate the candidate "unusual" events signaled by the previous step. This step is analogous to human participation in data mining to validate the unusual patterns output by the data mining algorithm.

Another advantage of this scheme is that we can use the likelihood values output from the HMM to rank the highlights. This avoids the simple heuristic in [4] that uses the duration of applause/cheering segments for ranking and summary length modulation. Figure 2 illustrates the combination of unsupervised and supervised learning approaches for highlight extraction.

4. Experimental Results

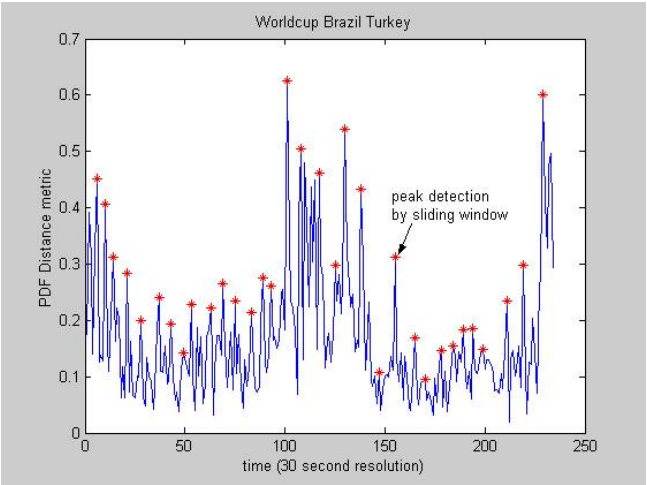
Training data was collected for each of the low-level sound classes from three and a half hours of MPEG video for three different sports namely baseball, soccer and golf. Twelve dimensional MFCC features extracted from every 30ms frame, were used to train a 10 component GMM for each sound class.

Once models are trained, input audio is divided into chunks of 1s segments. MFCC features are extracted from each frame in the 1s segment. Each frame is then classified as belonging to one of the sound classes using the maximum likelihood criterion. We use a majority voting scheme from all the frames to decide the sound class of the 1s segment.

After assigning audio labels to 1s segments, unsupervised label mining was performed to detect unusual events as shown in Figure 1. The value of large window (W_L) was chosen to be 4 minutes and the value of was small window (W_S) chosen to be 0.5 minutes. Kullback-Leibler distance metric was used to compare the distribution of audio labels within these chosen windows. Sliding window peak detection algorithm was used on the recorded distance values (Mw_s) to detect peaks. Figure 3 shows the peak detection results for two soccer games each of duration two hours.

After identifying candidate "interesting" events from the unsupervised label mining step, we use a Hidden Markov Model (HMM) to validate and rank the "interesting" events. The highlight HMM was trained from the labels extracted from DVD data which contains half an hour of professionally edited highlights of the 2002 soccer world cup. There were 81 "interesting" clips in this training data including mainly attempts at goal and goals. The number of states for the HMM was empirically chosen to be 4. After training, it was observed that one of the states corresponded to the Cheering label and the corresponding state had a high self-transition probability. This implies that the HMM was indeed modeling the occurrence of contiguous cheering segments. This kind of modeling is better than simply using the length of duration of cheering segments as in [4]. To illustrate this we include the corresponding precision-recall figures for the same soccer games in Table 2. The threshold based scheme in [4] to detect highlights, is based on a fixed notion of highlights and is less flexible. Here, we let the HMM learn from the label sequences of training data what a highlight is. Such a data driven approach does not have a fixed notion of highlights and may capture other audio cues apart from cheering/applause. It also opens the possibility of information fusion with other modalities.

Table 1 shows the performance in the two soccer games.



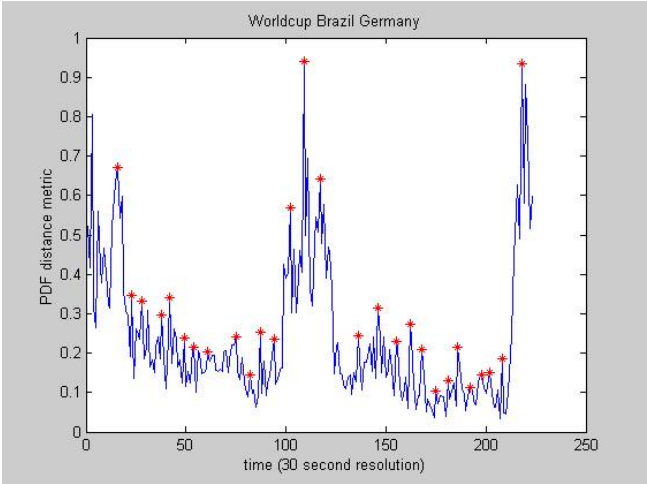


Figure 3: Unsupervised label mining results on soccer games.

	[A]	[B]	[C]	Precision	Recall
Game 1	34	22	32	40.74%	64.71%
Game1	34	6	8	42.86%	17.65%
Game 1	34	16	24	40%	47.06%
Game 1	34	11	20	35.48%	32.35%
Game 2	45	25	35	41.67%	55.55%
Game 2	45	5	0	100%	11.11%
Game 2	45	20	14	58.82%	44.44%
Game 2	45	9	4	69.23%	20%

Table 1: Soccer Game 1: World cup Brazil Germany; Soccer Game 2: World cup Brazil Turkey; [A]: Number of true highlight segments; [B]: Number of true highlight segments output by the algorithm; [C]: Number of False Alarms

	[A]	[B]	[C]	Precision	Recall
Game 1	34	22	59	27%	64.71%
Game 1	34	6	11	35%	17.65%
Game 1	34	16	39	29%	47.06%
Game 1	34	11	30	27%	32.35%
Game 2	45	25	44	36%	55.55%
Game 2	45	5	10	32.5%	11.11%
Game 2	45	20	30	40%	44.44%
Game 2	45	10	14	42%	20%

Table 2: Results for the same two games using the approach in [4].

In order to compare the current approach with the threshold based scheme in [4], we compare the precision values for the same recall values in two soccer games. For the current approach, different points on the precision-recall curve were generated by changing the likelihood threshold of the highlight HMM.

Note that the proposed approach outperforms the simple threshold based scheme in [4] for the recall values. The number of false alarms has been reduced by the inclusion

of the peak detection and a validation step in place of a threshold based highlight detection.

5. Conclusion

We have presented our latest improvement on the sports highlights extraction framework described in [4]. These improvements can be summarized as a combination of unsupervised label mining method and a supervised highlight model. In [4], we have relied on the correlation between the length of applause/cheering segments and highlights without explicit modeling of these highlights. In this paper, we use highlight models together with the unsupervised mining of "unusual" segments, to further improve the performance.

In the future, we will further investigate the fusion of audio and video domain features to model the highlights.

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] A. Divakaran, K. A. Peker, R. Radhakrishnan, Z. Xiong and R. Cabasson, "Video Summarization using MPEG-7 Motion Activity and Audio Descriptors", Video Mining, eds. A. Rosenfeld, D. Doermann, and D. DeMenthon, Kluwer Academic Publishers, 2003.

[4] Z.Xiong, R.Radhakrishnan, A.Divakaran, "Generation of Sports Highlights Using Motion Activity in Combination with a Common Audio Feature Extraction Framework", IEEE International Conference on Image Processing (ICIP), To Appear September 2003.

[5] B. L. Yeo, B. Liu, "Rapid scene analysis on compressed video", IEEE Trans. On Circuits and Systems for Video Technology, 5(6):533–544, Dec 1995.

[6] M. J. Swain, D. H. Ballard, "Color indexing", Int. J. Comput. Vision, 7:11-32. 1991.

[7] L.Xie, S.F.Chang, A.Divakaran, H.Sun, "Unsupervised Mining of Statistical Temporal Structures in Video" , Video Mining, eds. A. Rosenfeld, D. Doermann, and D. DeMenthon, Kluwer Academic Publishers, 2003.

[8] H.Pan, P.Van Beek, M.I.Sezan, "Detection of slow-motion replay segments in sports video for highlights generation", in Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing, 2001.

[9] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization", accepted for publication in IEEE Trans. Image Processing

[10] M.Xu, N.Maddage, C.Xu, M.Kankanhalli, Q.Tian, "Creating Audio Keywords for Event Detection in Soccer Video", Proc. of ICME 2003.

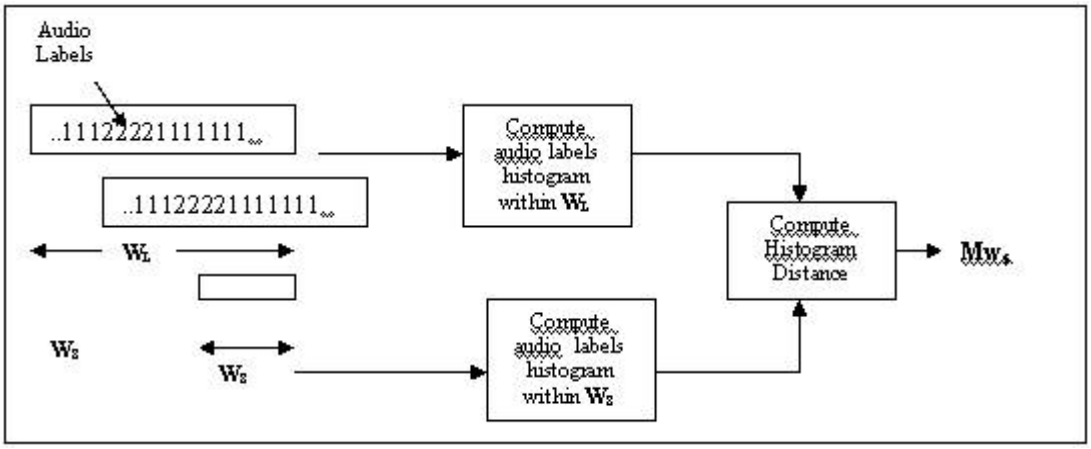


Figure 1: Unsupervised Label Mining Framework

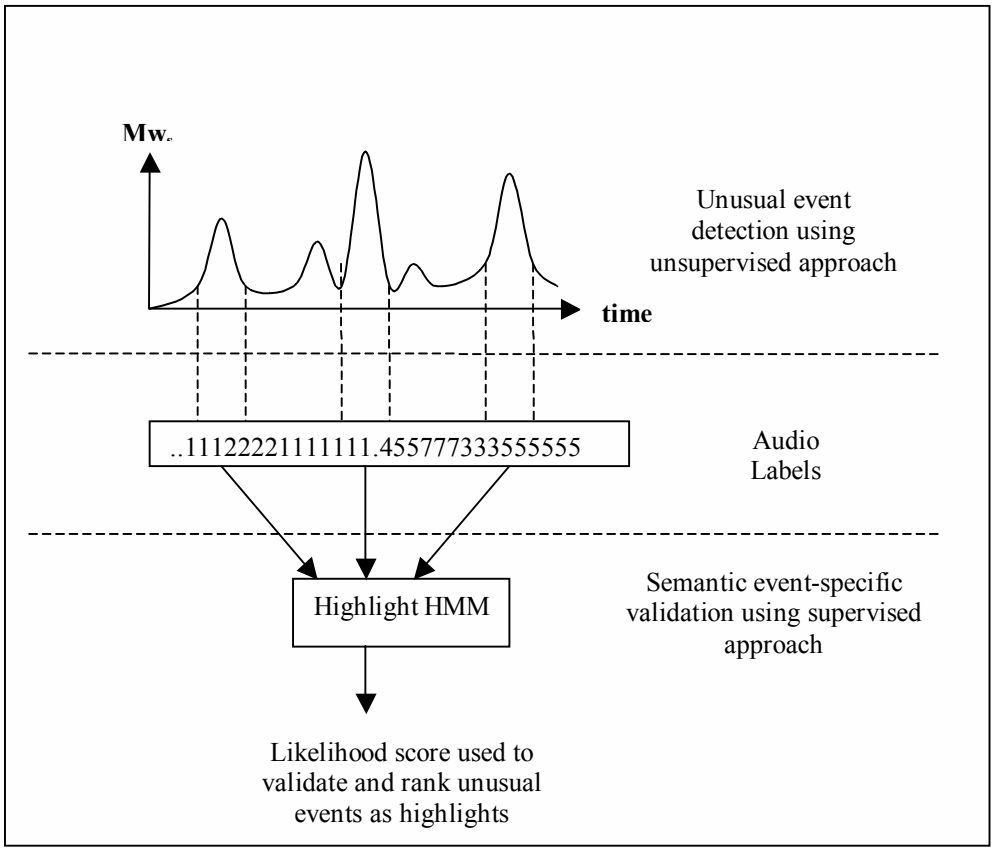


Figure 2: Combination of Unsupervised and Supervised learning for highlight extraction.