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

Audio-Assisted Video Browsing for DVD Recorders

Ajay Divakaran, Isao Otsuka, Regunathan Radhakrishnan, Kazuhiko Nakane, and Masaharu Ogawa

TR2004-139 December 2005

Abstract

We present an audio-assisted video browsing system for a Hard Disk Drive (HDD) enhanced DVD recorder. We focus on our sports highlights extraction based on audio classification. We have systematically established that sports highlights are indicated by the presence of audience reaction such as cheering, applause and the commentatorś excited speech. That enables us to develop a common highlights extraction technique, based on detection of audience reaction, for five different sports, viz. soccer, baseball, golf, sumo wrestling and horseracing. Our extraction accuracy is high. Furthermore, the percentage duration of the audience reaction gives us a simple importance measure for each of the highlights. We can then get a summary of any desired length by appropriately choosing a threshold for the importance measure. We process the AC-3 audio directory thereby enabling simple integration of our technique into our target platform.

IEEE PCM Conference 2004

Audio-assisted Video Browsing for DVD Recorders

Ajay Divakaran[1], Isao Otsuka[2], Regunathan Radhakrishnan[1], Kazuhiko Nakane[2], and Masaharu Ogawa[2]

[1] Mitsubishi Electric Research Laboratories, Cambridge, USA,
ajayd@merl.com regu@merl.com,
[2] Mitsubishi Electric Corporation, Kyoto, Japan

Abstract. We present an audio-assisted video browsing system for a Hard Disk Drive (HDD) enhanced DVD recorder. We focus on our sports highlights extraction based on audio classification. We have systematically established that sports highlights are indicated by the presence of audience reaction such as cheering, applause and the commentator's excited speech. That enables us to develop a common highlights extraction technique, based on detection of audience reaction, for five different sports, viz. soccer, baseball, golf, sumo wrestling and horseracing. Our extraction accuracy is high. Furthermore, the percentage duration of the audience reaction gives us a simple importance measure for each of the highlights. We can then get a summary of any desired length by appropriately choosing a threshold for the importance measure. We process the AC-3 audio directly thereby enabling simple integration of our technique into our target platform.

1 Introduction and Motivation

Our target application is the Personal Video Recorder (PVR) (See Nakane et al [2003]) . Current PVR's can store up to 200 hours of video content, and the storage capacity is expected to further grow in the future. There is therefore a great need for video browsing techniques that help the user skim through many hours of content quickly, as well as to select which part of the content to play in full. While video browsing and summarization (See for example Divakaran et al [2003]) has been an active area of research, realization of video summarization techniques on consumer video devices is an open challenge. First, there is no constraint on the variety of content genres that the consumer will record, so the summarization algorithms have to be applicable to a wide variety of content. Second, consumer video devices typically use partly-programmable custom integrated circuits that offer limited computational resources.

In this paper, we propose a common framework based on processing the audio to gauge audience reaction for extraction of highlights of five sports, Soccer, Baseball, Golf, Sumo wrestling and horse racing, which is also applicable to other sports in which audience reaction is indicative of an interesting event. Since the compressed audio coefficients are in the frequency domain, our audio

classification directly works on them thus obviating the computation required for conversion to the frequency domain. We thus take advantage of the target platform architecture to vastly reduce the computational complexity. The rest of the paper is organized as follows. In section 2, we describe the audio classification framework. In section 3, we describe the highlights extraction based on detection of audience reaction and the ranking of highlights and present some experimental results. In section 4, we describe the implementation of the proposed algorithm on the target platform. In section 5, we present our conclusions and possibilities for further research.

2 Audio Classification Framework

The following sound classes span almost all of the sounds in sports domain: Applause,"The commentator's Excited Speech combined with Cheering," " Cheering", Music, Speech & Speech with Music (See Xiong et al [2003]) . We collected training data from several hours of broadcast video data for each of these sound classes. We extract MDCT (Modified Discrete Cosine Transform) coefficients from the AC-3 encoded audio, at the rate of 32 frames per second, with each frame containing 256 coefficients. The 256 coefficient frame is considered to be our feature vector and is normalized using the audio energy in the last second of the clip. We reduce the dimension of the 256 dimensional feature vector through Principal Component Analysis (PCA), to 22. We trained Gaussian Mixture Models (GMMs) to model the distribution of features for each of the sound classes. The number of mixture components were found using the minimum description length principle so as to get high classification accuracy (See Xiong et al [2004]). Then, given a test clip, we extract the features for every frame and assign a class label corresponding to the sound class model for which the likelihood of the observed features is maximum. We illustrate our audio classification in Figure 1.

Our audio classification techniques are robust in the face of the extremely noisy ambience of broadcast sports video. Such robustness stems from the careful collection of the training data as well as the use of the Minimum Description Length principle while training the GMM's. However, the accuracy is clearly not perfect, and hence we have to compensate for the mis-classifications when we extract the highlights.

3 Sports Highlights Extraction

Our sports highlights extraction is guided by the intuition that interesting events lead to notable audience reaction in the form of applause or cheering combined with excited speech from the commentator. Furthermore, the more interesting the event, the longer the audience reaction to the event. For example, a goal scoring move in soccer would get an audience to cheer and the commentator to comment excitedly for a much longer time than would a mildly interesting moment such as a free kick. Therefore, the audience reaction provides a powerful indicator of interesting events or highlights across a wide range of sports. We

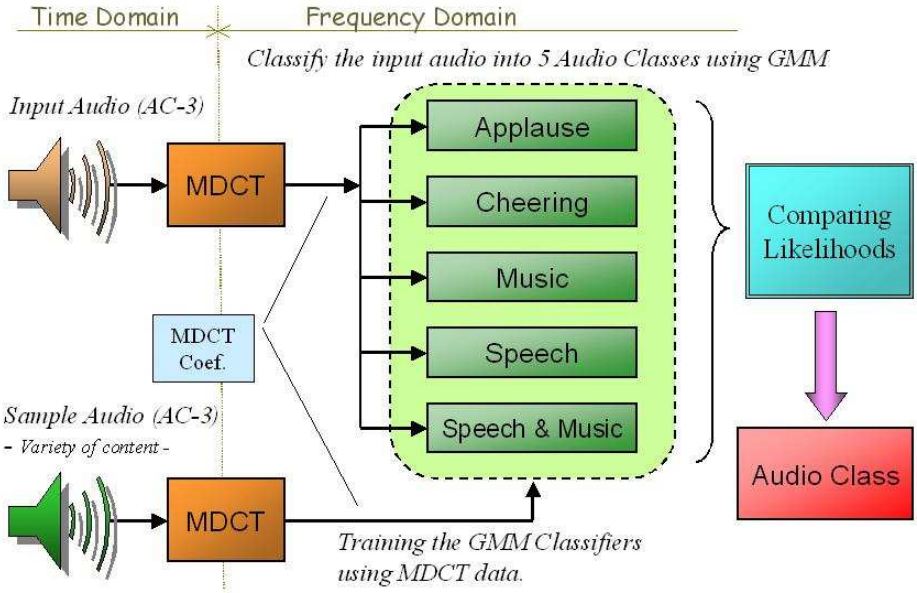

Fig. 1. Audio Classification Framework

thus have a genre-independent way to extract sports highlights as illustrated in Figure 2. We compute the percentage of the key audio class such as applause or excited speech and cheering in a sliding window of a certain pre-set length such as 10 seconds. Note that unlike our previous work (Xiong et al [2003]) we do not use the length of contiguous segments of unbroken audience reaction. We use the percentage instead to compensate for mis-classification that falsely breaks up long stretches of audience reaction. Furthermore, we weight the percentage of audience reaction in the window with the audio energy. We do so to eliminate false alarms caused by low energy audio that is mis-classified.

As shown in Figure 3, our sports highlights extraction gives rise to an importance measure for every time unit of the video sequence. We can then get a summary of a desired length by setting the threshold for the importance appropriately. Thus for example, if we want just the few most important highlights, we would increase the threshold until the right number of highlights are above the threshold.

We have tried our approach with a wide variety of sports content drawn from Soccer, Baseball, Golf, Sumo Wrestling and Horse-racing. We have got satisfactory accuracy with 30 odd games from Japanese and U.S. broadcast sports content. We illustrate a typical result with a Golf game in Figure 4. We determine the accuracy by first manually annotating notable events such as goals, home runs etc., and then checking to see if our automatic technique detects them. We find that we have very few misses when we aim for setting the

summary duration to roughly ten percent of the total duration of the content. However, note that since the threshold is variable, the user can adjust it to get better results. We believe the reason for the good performance is that audience reaction is a reliable indicator of how notable the event is. The audience reaction is in fact an indication of the consensus among the human observers of the event on how remarkable the event is. Therefore, if the audio classification works well, the highlights extraction is very successful. We also find that with sports such as golf that have low ambient sound because the spectators are relatively quiet,we get much better results than with sports such as soccer in which the audience is noisy throughout. However, the excited speech of the commentator is spectrally so distinctive, that it is detectable even when the ambience is noisy.

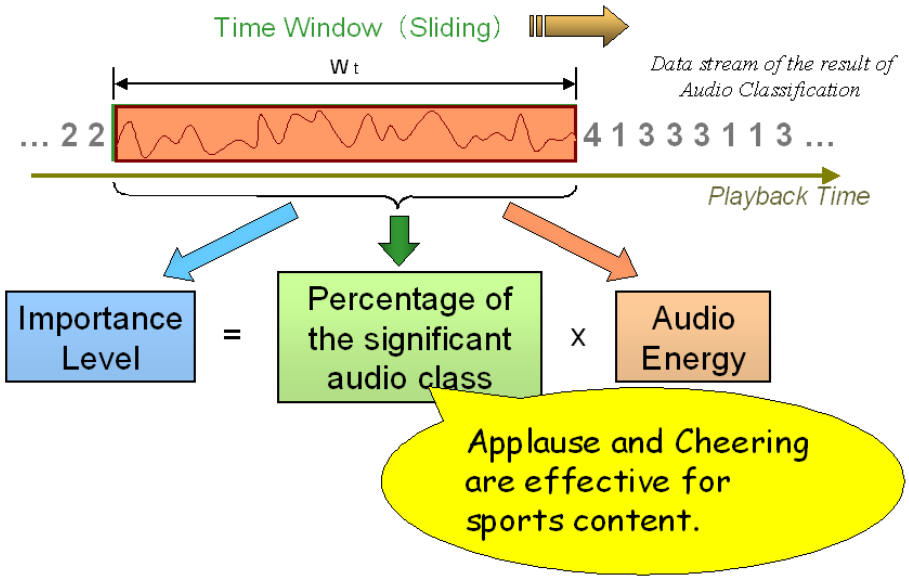

Fig. 2. Sports Highlights Extraction Framework

4 Realization on Target Platform

Since our sports highlights extraction uses the same essential algorithm for a wide variety of sports content, it requires only a simple enhancement of the target platform viz. the DVD recorder. We illustrate our proposed enhancement in Figure 5. Furthermore, our use of the Minimum Description Length principle while training GMM's gives rise to compact GMM's with as few mixture components as is feasible while maintaining accuracy. Notice also that our direct use of

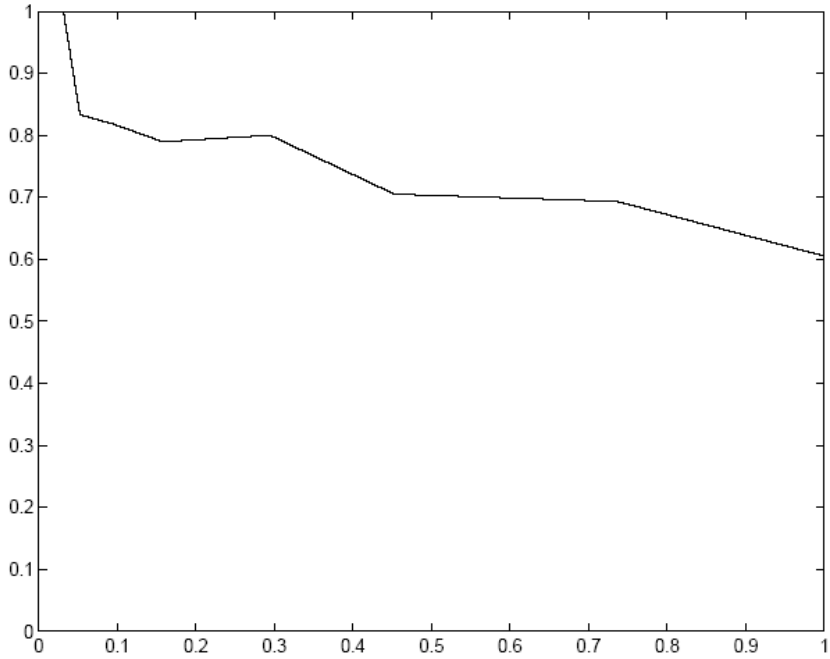

Fig. 3. Precision (y) -Recall (x) Plot for British Open Golf Game

the AC-3 coefficients makes our enhancement a simple and direct modification of the audio encoder.

The summarization meta-data is generated during the recording phase as can be seen in Figure 5. Since we use a sliding window, there is a delay equal to the length of the window after the recording is completed. However, such a delay is only a few seconds, hence the summarization meta-data is available to the user almost as soon as the recording is complete. The meta-data is written out to the HDD or to the DVD disc depending on the user preference. Our meta-data syntax is simple since it is merely an importance measure profile of the video as illustrated in Figure 3. It therefore takes up very little space and can be written in the private space of the DVD disc.

5 Conclusion

We presented a common highlights extraction algorithm for a wide variety of sports using audio detection of audience reaction to notable events, that has reasonable accuracy. We proposed an approach to realization of the algorithm

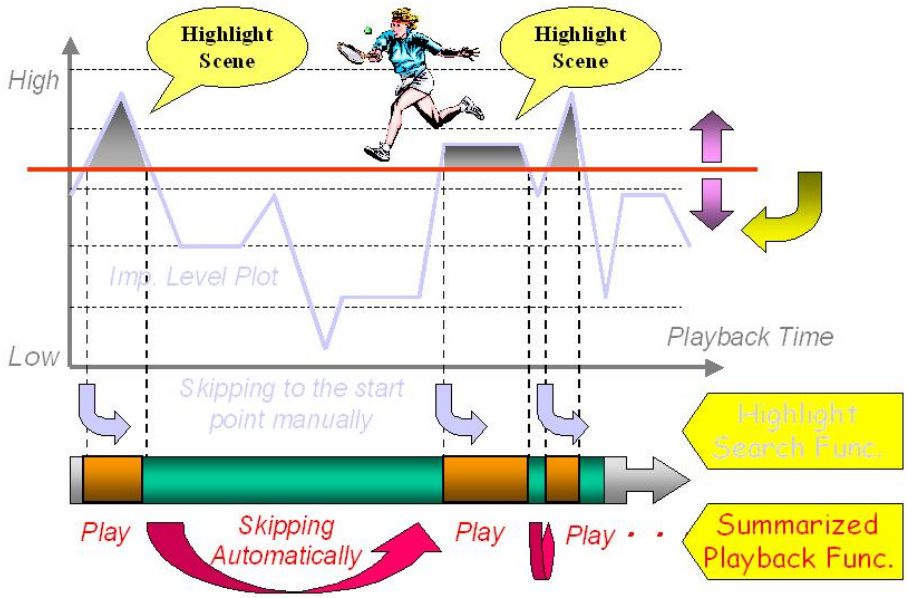

Fig. 4. Scalable Summarization

on our target platform. Note that our algorithm design from the outset ensured simple integration into the target platform. We will present a demonstration of our algorithm at the conference.

There are several avenues for further improvement of the system. First, we will explore improvement of the audio classification accuracy. Second, we will explore incorporation of visual cues into the system. We hope to eliminate false alarms as well as carry out a richer segmentation of the content. Third, we will extend our framework beyond sports video. Our emphasis has been on "infotainment" content such as sports and news because we believe that summarization is best suited for browsing for information. However, we now need to find out how to summarize other genres such as dramas and documentary movies, and how to accommodate our extension of the algorithms on our target platform. Fourth, we plan to explore the user interface. In our view, the user interface is the critical component of the system that can much more than compensate for lack of accuracy in detecting sports highlights. We need to develop a user interface that will work well in our target platform which is remote-control driven.

References

[2003] A. Divakaran, K. A. Peker, R. Radhakrishnan, Z. Xiong and R. Cabasson, *Video Summarization using MPEG-7 Motion Activity and Audio Descriptors,* in Video

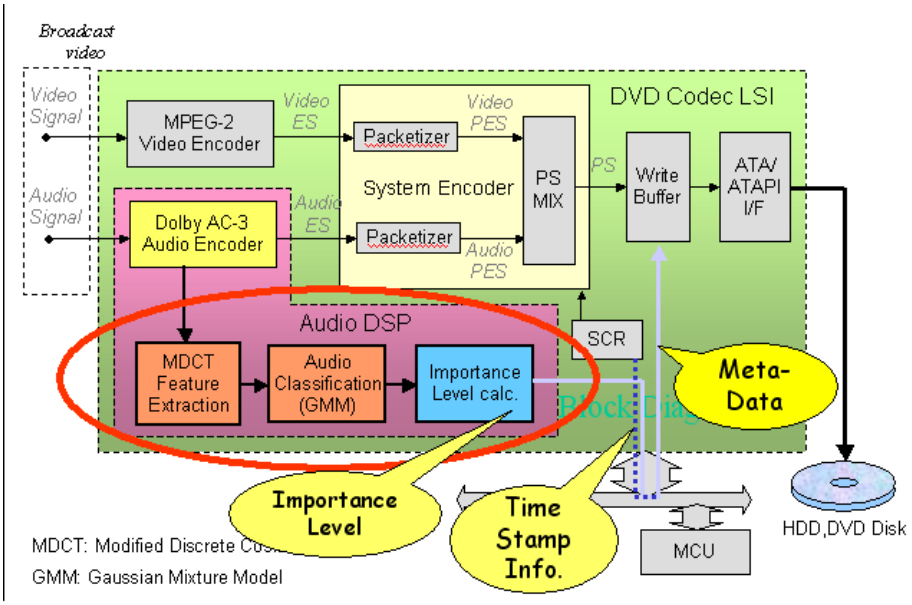

Fig. 5. Realization on Target Platform

Mining, A. Rosenfeld, D. Doermann, and D. DeMenthon, Eds., Kluwer Academic Publishers, 2003.

[2003] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T.S. Huang, *Audio Events Detection based Highlights Extraction from Baseball, Golf and Soccer Games in A Unified Framework,* ICASSP 2003, April 6-10, 2003.

[2003] K. Nakane, I. Otsuka, K. Esumi, T. Murakami and A. Divakaran, ,*A Content-based Browsing System for HDD and/or recordable DVD Personal Video Recorder,* IEEE Conference on Consumer Electronics (ICCE), 2003.

[2004] I. Otsuka, A. Divakaran, K. Nakane, and M. Ogawa,*HDD Enabled DVD Recorder System,* IPSJ Conference, Japan, March 2004.

[2004] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T.S. Huang,*Effective and Efficient Sport Highlights Extraction using the Minimum Description Length Criterion in Selecting GMM Structures,* ICME 2004, June 27-30, 2004.