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

A Highlight Scene Detection and Video Summarization System Using Audio Feature for a Personal Video Recorder

Isao Otsuka, Kazuhiko Nakane, Ajay Divakaran, Keiji Hatanaka, Masaharu Ogawa

TR2005-011 December 2005

Abstract

The Personal Video Recorder such as Recordable-DVD Recorder, Blu-ray disc Recorder and/or Hard Disc Recorder has become popular for a large volume storage device for video/audio content data and a browsing function that would quickly provide a desired scene to the user is required as an essential part of such a large capacity recording/playback system. We propose a highlight scene detection function by using only Áudiofeatures and realize a browsing function for the recorder that enables completely automatic detection of sports highlights.

ICCE 2004

A Highlight Scene Detection and Video Summarization System using Audio Feature for a Personal Video Recorder

Isao Otsuka[1)], Kazuhiko Nakane[1)], Ajay Divakaran[3)], Keiji Hatanaka[2)], and Masaharu Ogawa[1)]

***Abstract*—The Personal Video Recorder such as Recordable-DVD Recorder, Blu-ray Disc Recorder and/or Hard Disc Recorder has become popular for a large volume storage device for video/audio content data and a browsing function that would quickly provide a desired scene to the user is required as an essential part of such a large capacity recording/playback system.**

We propose a highlight scene detection function by using only 'Audio' features and realize a browsing function for the recorder that enables completely automatic detection of sports highlights.

I. INTRODUCTION

The Personal Video Recorder (PVR) such as Recordable-DVD, Blu-ray Disc Recorder and/or Hard Disc (HDD) Recorder has become popular for a storage device of large volume video/audio content [1].

A browsing function that would quickly provide a desired scene to the user is required as an essential part for such a large capacity recording system.

In our previous work [2], we proposed a video browsing system using a combination of motion based video summarization and topic-related metadata in the incoming video stream with audio assistance.

In the case of a sports program, highlight scenes elicit human reaction in form of the excited voice of the commentator and cheering by the audience. So, we propose a highlight scene detection function by using only 'Audio' features to detect human reaction and realize a browsing function for the recorder that enables completely automatic detection of sports highlight.

II. PROPOSED SYSTEM

A. System Configurations

A simplified block diagram of the investigated video browsing system in the recording phase is shown in Figure 1.

For example, the video and audio signals from a broadcast video are encoded using MPEG2 and AC-3, packetized, and stored onto a disc such as HDD, DVD, Blu-ray medium via buffer. The video encoder is entirely hard-wired but the audio encoder is based on a programmable DSP in general, therefore it is easy to add our proposed modification which includes 3 modules the 'MDCT feature extraction block', the 'Audio Classification block', and the 'Importance Calculation block' as shown in Figure 1.

1) Advanced Technology R&D Center, Mitsubishi Electric Corporation., Kyoto Japan
2) Kyoto Works, Mitsubishi Electric Corporation, Kyoto Japan.
3) Mitsubishi Electric Research Laboratories, Cambridge, MA

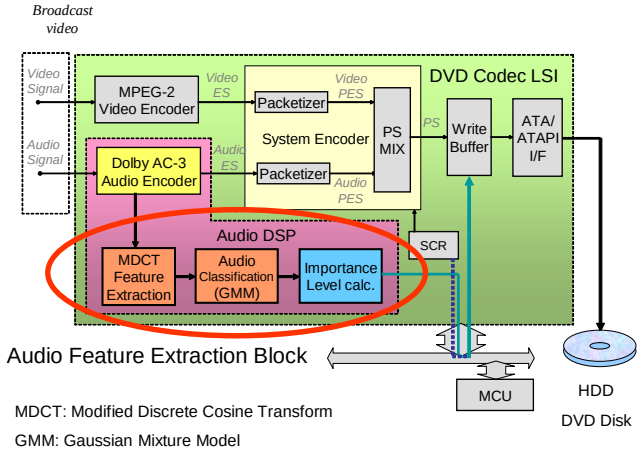

Fig. 1. Simplified Block Diagram for the Recording Phase

B. Importance Level Calculation

The Audio Classification block in the audio DSP classifies each audio segment as one of several classes (e.g., Applause, Cheering, Excited Speech, Normal Speech, Music, etc.) using low-complexity Gaussian Mixture Models (GMM)

We trained the GMM classifiers using MDCT coefficients from a variety of sports content. So the system classifies the input audio by comparing the likelihoods of the audio classes as shown in Figure 2. The reason why we use MDCT coefficients is that they are already available in the AC-3 encoding phase thus saving the computation needed to convert the time domain signal into the frequency domain.

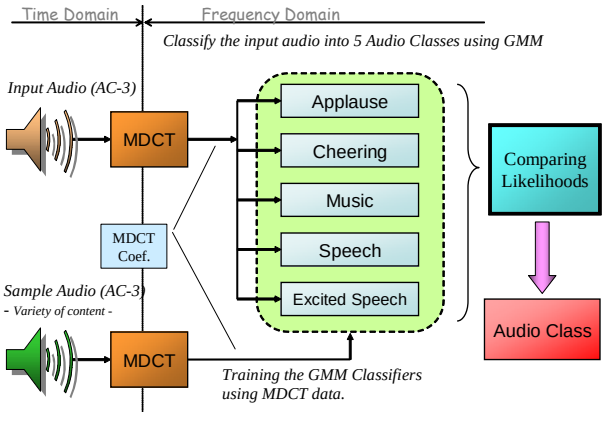

Fig. 2. Audio Class Recognition (GMM)

The Importance Calculation block calculates the percentage of the significant audio class in a time window of an audio classification data stream to get the importance level for each audio segment. In the case of sports content, using "Excited

Speech" as the significant class is effective.

The importance level and start/end time information (i.e. PTM) for each audio segment are stored onto the medium as a unique meta-data file.

The simple feature extraction allows all the meta-data generation to be done at the audio encoding phase in one pass.

The meta-data file is then written out in a separate directory in the medium as shown in the Figure 3.

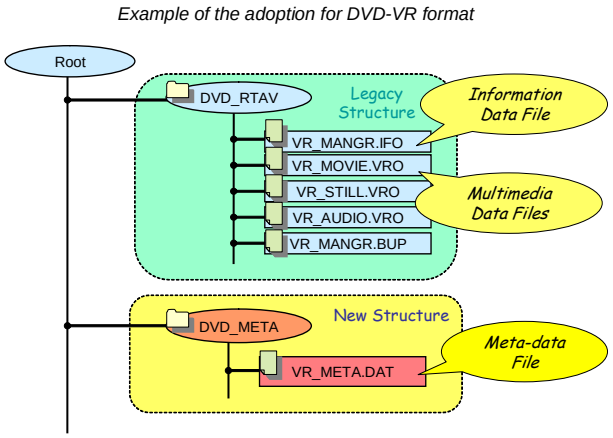

Fig. 3. D

C. *Highlight Search/Playing Back Function*

In the second pass, i.e. the playing back phase, the system takes the importance level plot by the reading only the meta-data file before the playing back of the target program on the disc and identifies the highlight scene, as shown in Figure 4.

From the user-interface point of view, a user can set and adjust the slice level. The portions with importance greater than the slice level chunks are identified as the highlights. So, skipping to the start position of the highlight scene manually is the function of 'Highlight Search', and skipping and playing back only the highlights is the function of 'Summarized Playback'. These unique functions give the user a powerful and useful way to browse the large volume content.

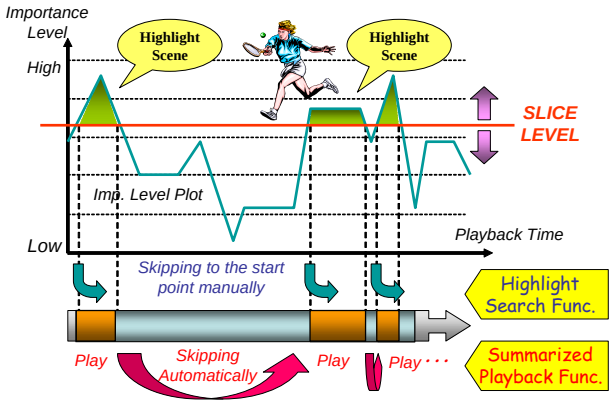

Fig. 4. Highlight Scene Detection and Playing back

III. PROTOTYPE MODEL DEVELOPMENT

We have developed a prototype model that performs the Summarization playback, and the Highlight Search function.

This prototype model was based on the current our DVD & HDD recorder product. It can display the importance level plot onto the screen and can adjust the slice level by manually as shown in the Figure 5.

Fig. 5. Prototype Model

The prototype model can generate the metadata, which includes an importance level for every second during the recording phase (real time), by DSP. We have tried over 50 sports videos, and obtained reasonable results. The use of the audio-based solution for the detecting highlight scenes has been shown to be effective especially for sports contents.

IV. CONCLUSION

This technology has been implemented in a personal computer simulation and a prototype model as well. The highlight scenes were detected from the contents of Baseball, Soccer, Horse Race, Sumo, and so on, and the accuracy was good enough for the practical use. We will give a demonstration at the conference.

We employed the audio-only solution for feature extraction, so we expect that the both the hardware and firmware would require simple enhancements to realize our approach on PVR.

ACKNOWLEDGMENT

The authors greatly appreciate the valuable contribution of Mr. Regunathan Radhakrishnan and Dr. Ziyou Xiong to studying and prototyping the system described in this paper.

REFERENCE

[1] K.Nakane, Y.Sato, Y.Kiyose, M.Shimamoto and M.Ogawa, "Development of Combined HDD and Recordable DVD Recorder-Player," *ICCE*, June 16-20, 2002, Los Angeles.

[2] K. Nakane, I.Otsuka, K.Esumi, A.Divakaran, and T. Murakami, "A Content-Based Browsing System for a HDD and/or Recordable-DVD Personal Video Recorder," *ICCE*, June 15-19, 2003, Los Angeles.