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

Highlight Scene Detection and Video Summarization for PVR-Enabled High-Definition Television Systems

S. Shipman, A. Divakaran, M. Flynn

TR2007-060 January 2007

Abstract

We describe an implementation of our previously developed highlight scene detection and video summarization system on a PVR-enabled high definition television system. The implementation poses significant challenges such as consuming minimal computational resources so as to avoid disrupting any of the existing functionalities. The target platform poses challenges such as the absence of a DSP which differ significantly from our previous PVR platform. Our implementation successfully addresses the aforementioned challenges by adapting techniques such as limited buffering and error concealment for dropped frames to the target platform.

IEEE International Conference on Consumer Electronics (ICCE

Highlight Scene Detection And Video Summarization for PVR-Enabled High-Definition Television Systems

Sam SHIPMAN[1], Ajay DIVAKARAN[1], and Mark FLYNN[2]
[1]Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA
[2]Mitsubishi Digital Electronics America (MDEA), Irvine, CA, USA

***Abstract*—We describe an implementation of our previously developed highlight scene detection and video summarization system, on a PVR-enabled high-definition television system. The implementation poses significant challenges such as consuming minimal computational resources so as to avoid disrupting any of the existing functionalities. The target platform poses challenges such as the absence of a DSP which differ significantly from our previous PVR platform. Our implementation successfully addresses the aforementioned challenges by adapting techniques such as limited buffering and error concealment for dropped frames to the target platform.**

I. INTRODUCTION

A previous paper [1] presented our highlight detection and video summarization system as implemented in a consumer product, a personal video recorder (PVR). This paper presents our prototype implementation of the same feature, targeted to a different consumer product: a high-definition television (HDTV) equipped with an internal digital personal video recorder. We compare the system architectures of the PVR and HDTV and discuss the implementation challenges posed by the HDTV and how they were met.

The highlight detection system analyzes a program while it is being recorded, to compute an importance-level metric at frequent intervals. These values are stored and subsequently used to allow the viewer to play only those portions of the program that exceed a specified importance level (i.e., the highlights). The importance-level computation uses the MDCT coefficients extracted from the AC-3 compressed audio. Gaussian Mixture Models trained with MDCT data typifying the audio classes of interest (e.g., excited speech, cheering) are used to classify each block of audio by determining the likelihood that the block belongs to each class. These likelihood values are used to compute importance levels for each second of audio, which are stored as meta-data associated with the program. Figure 1 diagrams the process as it occurs in the PVR product.

During playback, the user specifies an importance threshold, which determines the duration of the highlights to be shown (or vice versa). Figure 2 illustrates the highlight playback process. The process is explained in more detail in [1].

II. SYSTEM COMPARISON

The PVR product provides a coherent subset of the functionality provided by the HDTV product. Possible user interaction with the PVR is more constrained and so it is easier to characterize the CPU load. On the HDTV, CPU load varies more dynamically. The PVR has a slightly slower general purpose CPU than that of the HDTV, but the PVR also has an additional DSP that is used for the audio analysis for highlights detection.

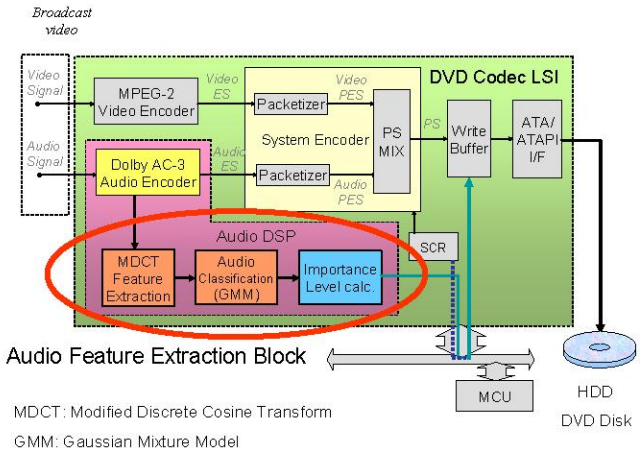

Fig. 1. PVR Implementation Block Diagram

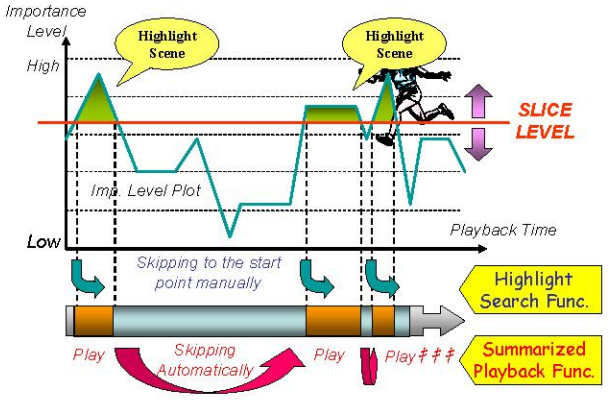

Fig. 2. Highlight Playback

In the PVR product, all content to be recorded is analog. We have access to the audio encoder implementation and can extract audio features during the encoding process. In the HDTV product, digital content is recorded directly, so the audio analysis software must perform a partial decoding of the content in order to access the audio features.

Because the HDTV has more functionality and must accommodate more varied user interaction, more physical main memory is provided. However, the HDTV platform is at

1-4244-0763-X/07/$20.00 ©2007 IEEE

a disadvantage for storing the highlight playback meta-data, because the PVR uses IEEE 1394 Audio/Video Control (AV/C) discs [2]. Since AV/C discs provide only a high-level interface for recording and playback of audio and video content, they do not provide direct access to the file system, so the meta-data cannot be stored there.

III. Implementation

It is a basic requirement that our implementation have as little impact on the structure, function, and performance of the existing system as possible.

The target platform lacks an additional DSP, so all computation for highlights detection must be performed on the platform's general purpose CPU. A substantial fraction of the CPU's capacity is already used in support of decoding and for other functions. The requirement to partially decode the compressed audio adds more computation, relative to that required for the PVR.

The straightforward approach would be to assign the highlights detection computation to a thread running at an appropriate scheduling priority and measure the total CPU usage, optimizing the computation until it runs within the available CPU capacity. However, because of the dynamically varying CPU load on the HDTV system, it is difficult to determine this *a priori*. If the computation could run completely asynchronously at its own rate, that might not matter. But the rate at which the computation must run is determined by the rate of incoming audio frames.

Given the larger amount of main memory available on the target platform, an alternative approach would be to buffer the compressed audio during recording, and process it asynchronously, finishing later than the end of recording if necessary. But it would be difficult to guarantee that sufficient memory would be available in all cases, given that memory can be dynamically allocated and programs to be recorded can be arbitrarily long (up to the available space on the recording medium). However, buffering is still useful for tolerating transient overloads.

Another approach would be to record the entire program first, and then read it back in and perform highlight detection. The AV/C disk complicates this approach because of the limited number of simultaneous readers supported and the low input rate (the real-time playback rate). Also the user may wish to use highlights playback immediately after (or even during!) recording, so this approach is not useful.

The approach we take is to run tests to ensure that highlights detection can run under typical load conditions, to use limited buffering to tolerate transient overloads, and to recover from dropped frames by repeating the meta-data generated from the previously computed frame.

One approach to further optimization of the highlights detection computation, suggested by the larger main memory available on the HDTV platform, is to use pre-computed tables to speed up the highlights detection computation. The tables would be statically allocated at load time. We are still implementing this optimization and the results will be presented at the conference. This optimization is not always successful--sometimes the system's cache behavior under load can dictate that performing the computation out of registers (if possible) is actually faster. The result will be of interest in either case.

Since the HDTV implementation cannot write meta-data to the AV/C disc, we must find another alternative, without adding hardware to the platform, if possible. Another form of non-volatile memory is available: the on-board flash memory. This memory is small and slow. We adapt to these conditions by adopting a compact meta-data representation and storing it in main memory until the recording is completed, then writing it out to flash.

IV. Conclusion

Our prototype implementation of highlight scene detection for the HDTV platform generates exactly the same importance levels as our PC-based reference implementation when running stand-alone on the target platform, even though the PC implementation has the benefit of much larger computational resources. We are still fine-tuning the implementation and the user interface is in development. Differences in architecture between the HDTV platform and the PVR platform on which our previous implementation was built resulted in different implementation challenges and trade-offs. Because the implementation requires no additional hardware, it incurs no additional manufacturing cost.

V. Acknowledgment

The authors greatly appreciate the valuable contribution of Mr. Atul Batra, formerly of MDEA and Dr. Kent Wittenburg of MERL for their consistent encouragement as well as the encouragement of Isao Otsuka and other colleagues at Kyoto Works, Mitsubishi Electric Corporation. We also wish to acknowledge the contributions of Regunathan Radhakrishnan, formerly of MERL and Daniel Ellis, Terence Pu, Brian Peterson, Mike Harvill, Polly Stecyk, Brian Maxson, and Peter Mortensen of MDEA.

References

[1] Otsuka, I.; Nakane, K.; Divakaran, A.; Hatanaka, K.; Ogawa, M., "A Highlight Scene Detection and Video Summarization System Using 'Audio' Feature for a Personal Video Recorder", *IEEE International Conference on Consumer Electronics (ICCE), Digest of Technical Papers*, pp. 223-224, January 2005.

[2] 1394 Trade Association Document 2002001, AV/C Disc Subunit General Specification 1.2, September 13, 2002.