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

MPEG-7 Meta-Data Enhanced Encoder System for Embedded Systems

Asai, K.; Nishikawa, H.; Kudo, D.; Divakaran, A.

TR2004-009 March 2004

Abstract

We describe a MPEG-7 Meta-Data enhanced Audio-Visual Encoder system that targets DVD recorders. We extract features in the compressed domain with both video and audio, which allows us to add the meta-data extraction without altering the hardware architecture of the encoder core. Our feature extraction algorithms are simple, and thus implementable through a simple combination of software and hardware on the integrated DVD chip. The primary application of the meta-data is video summarization, which enables rapid browsing of stored video by the end user. The simplicity of our summarization and feature extraction algorithms enables incorporation of the powerful functionality of smart content navigation through content summarization, into the DVD recorder at a low cost.

Publication History:

1. First printing, TR-2004-009, March 2004

MPEG-7 meta-data enhanced encoder system for embedded systems

Kohtaro Asai[1], Hirofumi Nishikawa[1], Daiki Kudo[1], Ajay Divakaran[2]
[1]Mitsubishi Electric – Information Technology Research Center, Ofuna, Kamakura, Japan
[2]Mitsubishi Electric Research Laboratories, Cambridge, MA 02139

ABSTRACT

We describe a MPEG-7 Meta-Data enhanced Audio-Visual Encoder system that targets DVD recorders. We extract features in the compressed domain with both video and audio, which allows us to add the meta-data extraction without altering the hardware architecture of the encoder core. Our feature extraction algorithms are simple, and thus implementable through a simple combination of software and hardware on the integrated DVD chip. The primary application of the meta-data is video summarization, which enables rapid browsing of stored video by the end user. The simplicity of our summarization and feature extraction algorithms enables incorporation of the powerful functionality of smart content navigation through content summarization, into the DVD recorder at a low cost.

Keywords: MPEG-7, DVD Recorders, MPEG-2, Compressed Domain Feature Extraction

1. INTRODUCTION

Personal Video Recorders (PVR) are rapidly gaining favor in the marketplace since they enable the consumer to record desired content and then browse it. So far the only content navigation that PVR's have provided has relied on information from the Electronic Programming guide. Once a desired program has been located, there are no tools to browse it rapidly other than using conventional trick play. Video Summarization addresses this need by presenting compact versions of the content that also work as collections of entry points into the content. For instance, the video summary of a news program can just be a collection of the starting frames of all the stories. Video Summarization is an active field of research, and almost all reported work relies on video analysis based on all or some of the available cues.

In this paper, we apply our video summarization to a Hard Disk Drive (HDD) enhanced DVD recorder system. We propose enhancements to the existing architecture so as to allow rapid browsing of the stored content. We use audio and video feature extraction in the compressed domain, which allows us to implement the summarization algorithms through a simple combination of hardware and software on our target platform at a low cost. The rest of the paper is organized as follows. Section 2 reviews our past video summarization work. Section 3 discusses challenges posed by the target platform. Section 4 describes our proposed system, and Section 5 presents our conclusions.

2. VIDEO SUMMARIZATION WITH AUDIO AND VIDEO DESCRIPTORS

In our previous work [1], we describe a video summarization technique based on sampling in the "Cumulative Motion Activity" space. We have found empirically that the intensity of motion activity in a video shot is a direct indication of its summarizability. Therefore segments with high motion should require more key-frames for summarization than do segments with low motion. The cumulative motion activity space warps the time axis in such a way that low motion activity segments are much shorter than are high motion segments. Thus uniform sampling of the cumulative motion activity gives more keyframes to the high motion segments than to the low motion segments. We find that the motion activity of a video segment can be easily computed using either the average or the standard deviation of the motion vectors, which can be readily extracted from the compressed bit stream. We have thus devised a scheme to quickly extract key-frames from a video sequence. Furthermore, we have devised a simple MPEG-7 threshold based method to find out how many key-frames a given video shot requires [1]. The video summary is then obtained by merely concatenating the extracted key-frames.

The above scheme works best when the semantic segment boundaries of the content are known. We turn to the audio stream to compute the semantic boundaries of the content. We use simple Gaussian Mixture Models (GMM's) for audio classification that use the MDCT coefficients from the AC-3 stream. We also carry out speaker change detection using the MDCT coefficients [4]. For news content for instance, knowing the speaker transitions and finding the principal speakers helps find the story boundaries (See Figure 2). We are thus able to summarize the content in two stages. In the first stage, we find the semantic segment boundaries, which allows us to skim the video by skipping from segment boundary to segment boundary. We then use the baseline motion based video summarization to summarize the individual segment of interest. In the case of sports video, we use patterns of motion and audio that are associated with highlights to detect interesting events. Note that a summary is finally in the form of time stamps that mark the beginning and end of video segments that constitute the summary.

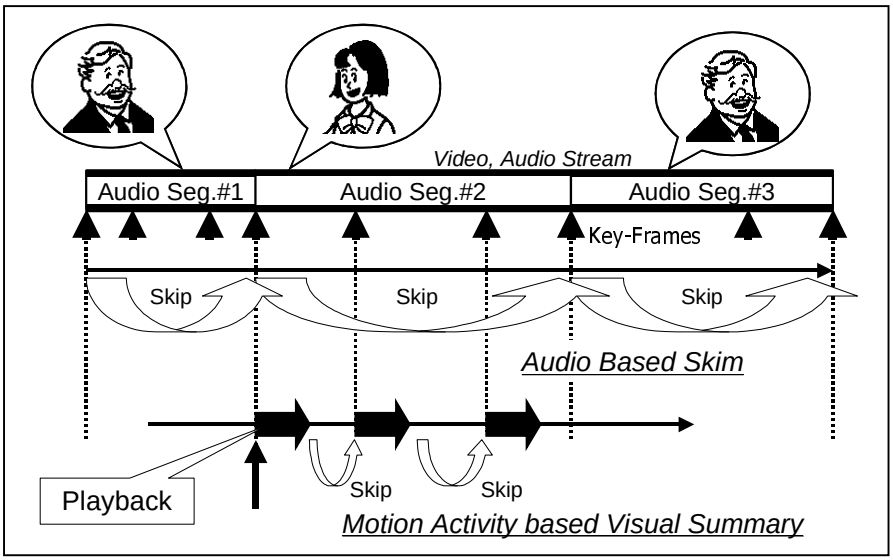

Figure 1: Audio Assisted Video Browsing

3. REQUIREMENTS OF DVD RECORDER SYSTEM

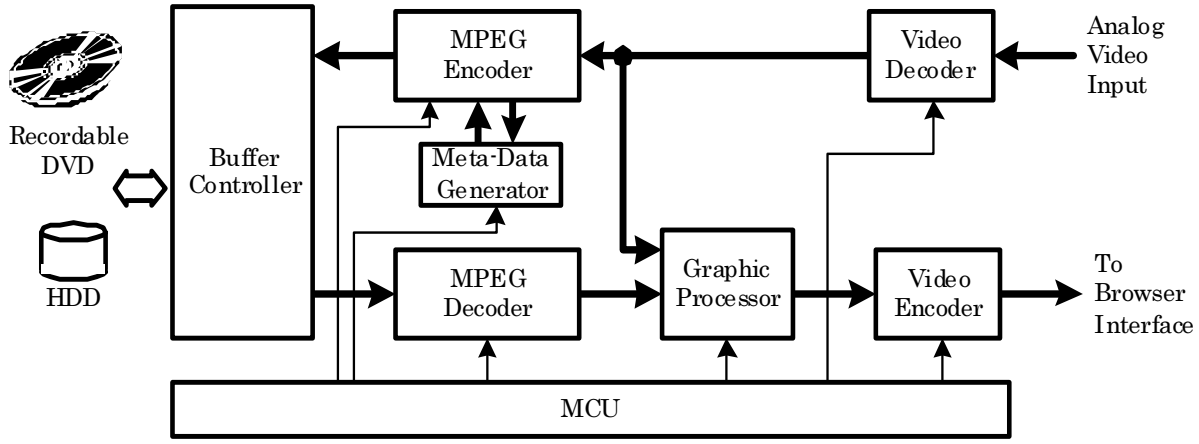

Figure 2: DVD Recorder System

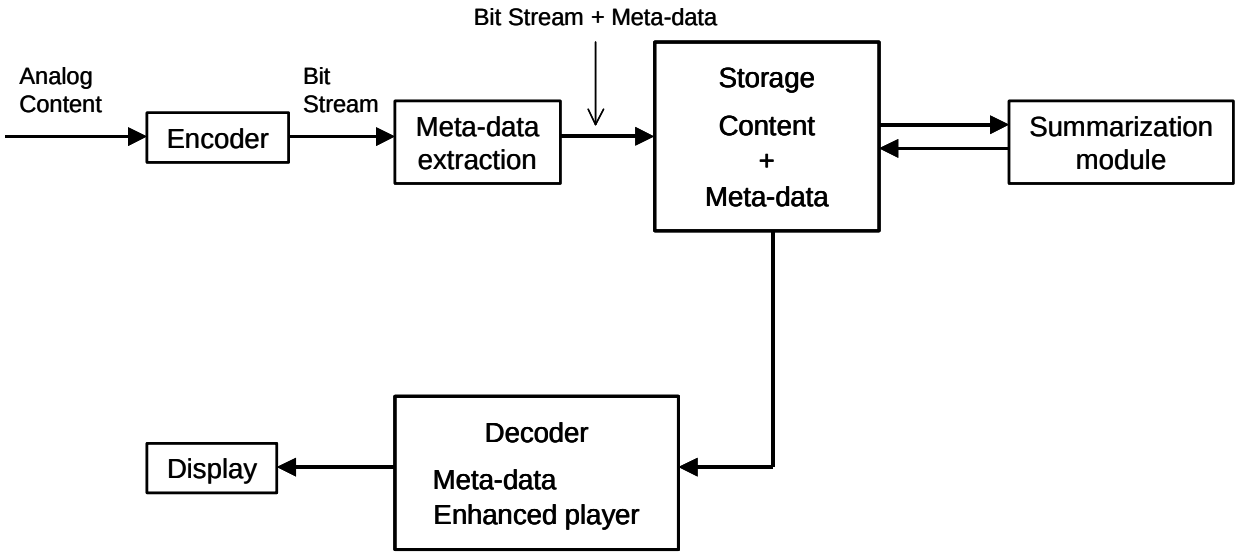

Figure 3: **Schematic Diagram of Meta-data Enhanced DVD Recorder.**

We illustrate a typical HDD enabled DVD recorder system in Figure 2, with a proposed meta-data generator. Let us enumerate the requirements placed on the meta-data generation:

1. Computational Simplicity
2. Easy integration with the existing platform
3. Fast meta-data generation with least or no waiting time for the user
4. Accuracy
5. Good user interface

Since the target is a consumer product, requirements 1 and 2 stem from a desire to minimize the cost impact of the meta-data enhancement. However, requirements 3, 4 and 5 stem from the end user experience. The meta-data generation has to be quick and has to be managed in such a way that there is little perceived delay between the recording of the content and its browsing. The accuracy is an obvious requirement since false alarms or misses will detract from the user experience. Finally, in all browsing systems, the user interface is a significant and complementary contributor to the success of the system. The user interface is outside the scope of this paper.

4. PROPOSED SYSTEM

. Our proposed video browsing system addresses the above requirements as follows:It relies on feature extraction in the compressed audio and video domain thus avoiding computationally costly inverse DCT's. It therefore meets the requirement of computational simplicity. It meets the easy integration requirement since feature extraction from the compressed domain can be achieved through either straightforward upgrading of the firmware, or through simple addition of custom hardware. Since the extraction is the compressed domain, the internal parts of the video codec need not be even accessed let alone modified. Thus our proposed meta-data generation fits easily into existing MPEG-2 codec based DVD recorder systems. As such codecs become more sophisticated, it will be possible to implement the meta-data generation purely through firmware, which will make the upgrading and integration process exceedingly simple and manageable.

While the meta-data generation is fast, it needs to be managed carefully so the waiting time for the user is kept at a low level. Since there are many possibilities for when the feature extraction would be carried out, we have to consider all the different possible scenarios and choose the most suitable. The most important decision to make is to decide whether to

put the summarization meta-data generation at the encoder or at the decoder. Placing it at the decoder places the computational burden on the player. Since DVD players are extremely low cost and mature products, it would not be reasonable to expect the decoder to be computationally strong enough to support the meta-data generation. Furthermore, it would increase the waiting time for the end user. We thus decide to generate the meta-data as the system is recording the content. This results in a small delay between the end of the recording and the end of the meta-data generation. The delay is 12 seconds or less, and is hence not especially noticeable. We could also use certain partial instantiations of the interface to mask the delay. In other words, we place the burden of summarization meta-data generation on the encoder as illustrated in Figure 3. Once this decision is made, the rest of the design decisions are straightforward. The player then makes use of the key-frame time stamps contained in the final meta-data. Note that such meta-data is negligible in size compared to the high volumes of content stored in a DVD or HDD. Also note that the entire process is the same regardless of whether the HDD or the DVD is used, thus allowing a seamless transition from and to the DVD. It also implies that we can use the same architecture for a generic PVR.

We have thus achieved a simple summarization enhancement based on audio-visual feature extraction in the compressed domain of a DVD recorder.

5. CONCLUSIONS & FUTURE WORK

We described a simple summarization meta-data enhanced DVD recorder system and the underlying design issues. We find that our system is simple and adds a small overhead to the encoder part of the MPEG-2 codec in exchange for a significant enhancement of the browsing functionality. We will present a demonstration at the conference.

In future work, we will target more sophisticated MPEG-2 codecs as well as further refine our audio-visual feature extraction based techniques to improve both simplicity and accuracy.

6.REFERENCES

1. 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.
2. http://www.merl.com/papers/TR2003-34/ Video Summarization using MPEG-7 Motion Activity and Audio Descriptors.
3. K. Nakane, I. Otsuka, K. Esumi, M. Ogawa, A. Divakaran, "A content-based browsing system for HDD and/or recordable DVD personal video recorder," *To appear in IEEE Transactions on Consumer Electronics*.
4. R. Radhakrishnan, Z. Xiong, B. Raj, A. Divakaran, "Effectiveness of mid-level feature representation for semantic boundary detection in news video," *Proc. SPIE Conf. Internet Multimedia Management Systems IV*, Orlando, FL, Sept 2003. (http://www.merl.com/papers/docs/TR2003-119.pdf)

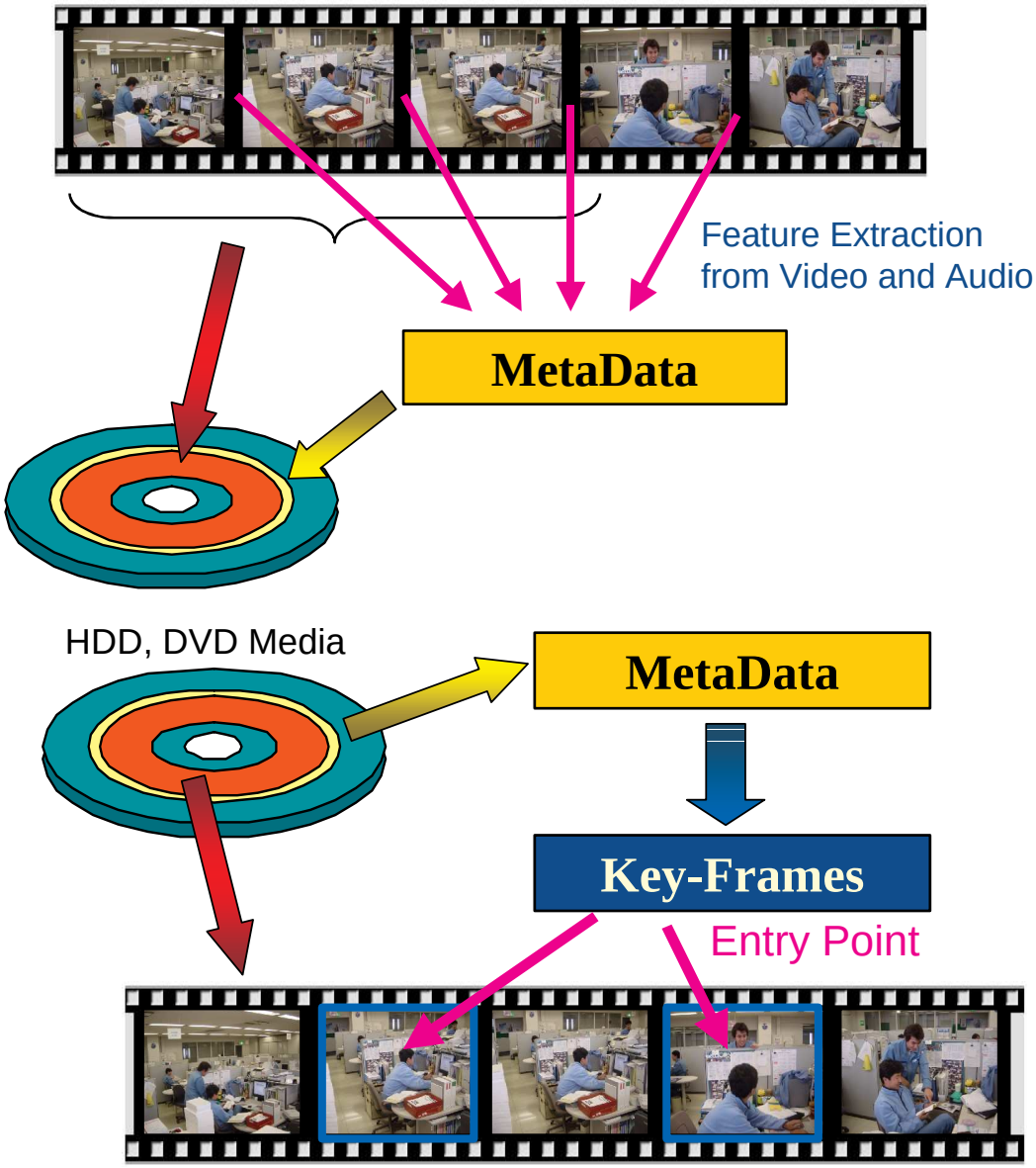

Figure 4: Recording onto DVD/HDD and Playback