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

Detection of Music Segment Boundaries using Audio-Visual Features for a Personal Video Recorder

I. Otsuka, H. suginohara, Y. Kushunoki, A. Divakaran

TR2007-037   February 2007

Abstract

We have extended our sports video browsing framework for personal video recorders, such as recordable-DVD recorders, blu-ray disc recorders and/or hard disc recorders, to music segment detection. Our extension to Japanese broadcast music video programs consists of detecting audio segment boundaries such as conversations with guests followed by music/song etc. Our proposed system first identifies the music/song scenes using audio analysis, and then adjusts the start/end position by detecting video shot changes, so as to achieve accurate detection of the music segment thus enabling rapid browsing. Our preliminary results indicate that our audio-only summarization with scene change support works well for music video content. We can therefore integrate the enhancement into our product at a low computational cost.

IEEE Transactions on Consumer Electronics

This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

Copyright © Mitsubishi Electric Research Laboratories, Inc., 2007
201 Broadway, Cambridge, Massachusetts 02139

Detection of Music Segment Boundaries using Audio-Visual Features for a Personal Video Recorder

Isao Otsuka, Hidetsugu Suginohara, Yoshiaki Kusunoki, and Ajay Divakaran

Abstract — *We have extended our Sports Video Browsing framework for Personal Video Recorders, such as Recordable-DVD Recorders, Blu-ray Disc Recorders and/or Hard Disc Recorders, to music segment detection. Our extension to Japanese broadcast music video programs consists of detecting audio segment boundaries such as conversations with guests followed by music/song etc. Our proposed system first identifies the music/song scenes using audio analysis, and then adjusts the start/end position by detecting video shot changes, so as to achieve accurate detection of the music segment thus enabling rapid browsing. Our preliminary results indicate that our audio-only summarization with scene change support works well for music video content. We can therefore integrate the enhancement into our product at a low computational cost.*

Index Terms — Video Summarization, HDD and DVD Hybrid Recorders, Music Detection, Sports Highlights Extraction.

I. INTRODUCTION

Commercially available Personal Video Recorders (PVR) such as HDD-enabled DVD [1] or Blu-ray Recorders, can already record several hundred hours of standard definition content and over fifty hours of High-Definition content. As storage capacities grow rapidly every year, it will be essential to develop rapid browsing technology that satisfies the consumer's need to get to a desired video segment quickly. In our previous work [2], we proposed a video browsing system using audio to detect sports highlights by identifying segments with a mixture of the commentator's excited speech and cheering, and also proposed to extend our strategy to music content by identifying music periods. [3]

In this paper, we describe a combination of three methods for identifying music/songs as a 'segment' with high accuracy. First, we use audio classification by using the Gaussian Mixture Models. Second, we detect the difference of audio energy between Right and Left channels, to establish the onset/end of the music. Third, we refine the start/end position of the music using scene change information.

II. PROPOSED SYSTEM FOR MUSIC VIDEO SUMMARIZATION

A. Application Framework

The proposed music video browsing and summarizing system is illustrated in Figure 1.

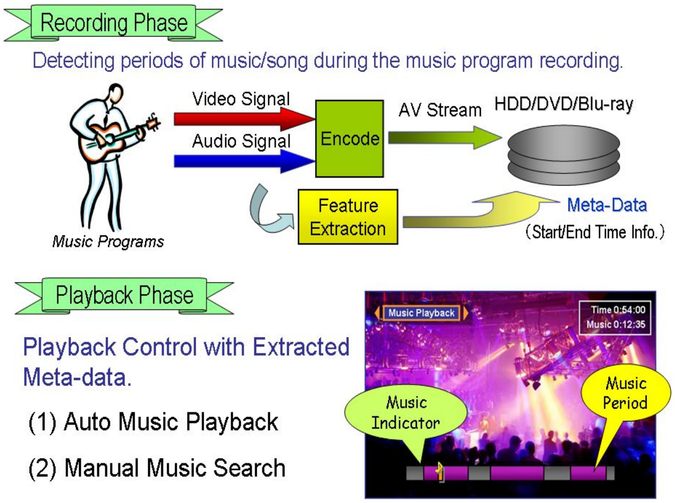


Fig. 1. Basic Concept of the Proposed Music Video Summarization

The basic objective of our proposal system is to find music or song segments from the entire recorded music program.

In the video recording phase, the multimedia data such as video and audio signals are encoded to digital AV data like a MPEG-2 video, Dolby AC-3 audio, and then stored onto a HDD/DVD.

Our proposed system analyzes audio and video features in real time. Then, based on the extracted features, the system identifies each music scene as a 'segment' in the recorded content. The detected start and end position (i.e. presentation time map) of music segments are stored onto the HDD/DVD as a Meta-data.

In the video playback phase, our proposed system reads out AV data with Meta-data from the disc. The detected start/end positions of music segments that are obtained from Meta-data can be plotted onto a graphical interface as shown in Fig.1. Thus we can propose the following functions when the system correctly detects the music periods; Skipping to the start or end position of music segment manually is the function of Music Search, and skipping and playing back only the music segments automatically is Auto Music Playback. The system allows jumping to end of music segments for the user who wants to watch only interviews or conversations with guest.

Isao Otsuka, Hidetsugu Suginohara, Yoshiaki Kusunoki are with the Advanced Technology R&D Center, Mitsubishi Electric Corporation, Kyoto, Japan (e-mail: Otsuka.Isao@cw.MitsubishiElectric.co.jp, Suginohara.Hidetsugu @dy.MitsubishiElectric.co.jp, Kusunoki.Yoshiaki@aj.MitsubishiElectric.co.jp).

Ajay Divakaran is with Mitsubishi Electric Research Laboratories, Cambridge, USA (e-mail: ajayd@merl.com).

Manuscript received January 6, 2007 0098 3063/07/$20.00 © 2007 IEEE

B. System Configuration

A simplified block diagram of the investigated music segment detection system is shown in Figure 2.

For example in the recording phase, the video and audio signals from analog broadcast video are encoded using MPEG2 video and AC-3 audio, packetized, and stored onto a disc such as HDD, DVD, Blu-ray Disc medium via buffer. The proposed system has 3 methods for identifying music/song segments, the 'Music Label Method', the 'Stereo Difference Method', the 'Scene Change Method' as shown in Figure 2.

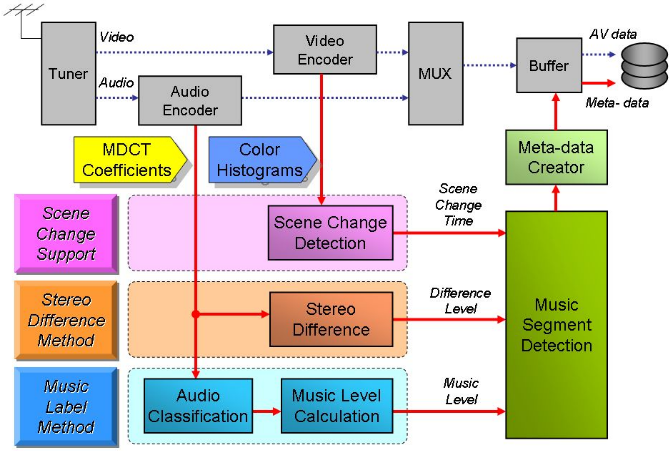


Fig. 2. Simple Block Diagram of the Music Period Detection

The 'Music Label Method' calculates the 'Music Level' by using MDCT coefficients, which are readily available from the AC-3 Audio Encoder. The method consists of the 'Audio Classification' block and the 'Music Level Calculation' block as shown in Figure 2.

The 'Stereo Difference Method' calculates 'Difference Level' by using the MDCT coefficients of Left and Right channels.

The 'Scene Change Method' calculates scene change times by using YUV color histograms from the MPEG-2 video encoder.

The results of the three methods are combined in the 'Music Segment Detection' block, and then the segmentation information is stored as Meta-data.

III. Music Segment Detection

For identifying music/song as a 'segment' with high accuracy, we propose to combine three methods as followings.

A. Music Label Method

The Audio Classification block classifies the input audio into one of 2 classes (Music or Talk) using low-complexity Gaussian Mixture Models (GMM) as shown in Figure 3. Note that the 'Talk' class covers non-music scenes such as interviews and conversations with a guest, and so on.

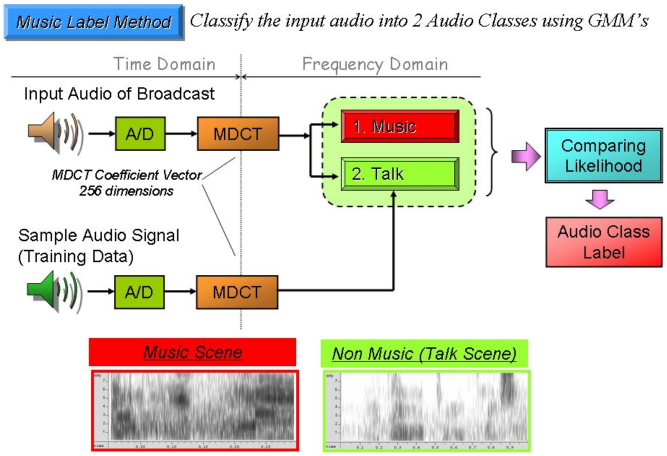


Fig. 3. Music Level Calculation with GMM

We trained the GMM classifiers using MDCT coefficients from a variety of broadcast music programs. So the system extracts the features for every frame and classifies a class label corresponding to the audio class model for which the likelihood of the observed features is the maximum.

The Music Level Calculation block computes the 'Music Level' which we define as the percentage of chunks classified as Music, for each second. An example of the Music Level plot is shown in Figure 4.

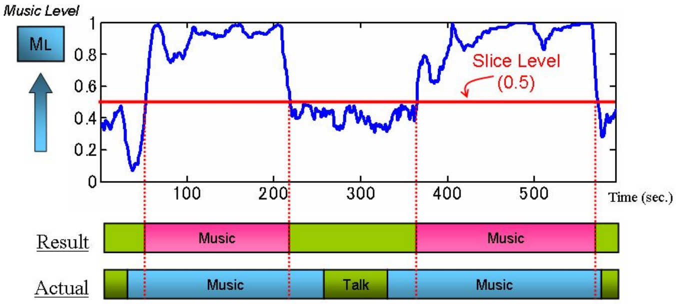


Fig. 4. Example of the Music Level Plot

Here is an example of an actual broadcast program which consists of music and talk scenes. The segments with music level above the slice level are determined to be the music scenes. As the figure shows, we lost some music parts, especially at the start and end of music periods. The reason is that the audio classifier cannot identify the music periods correctly since there are very quiet or intermittent melodies in the portions.

So we can say that this method finds music/song scenes with high reliability but can only vaguely identify the start/end positions of music segments

B. Stereo Difference Method

Usually, a music TV program is broadcast in STEREO mode. We find with such content that the difference between Left and Right channels is large in the frequency domain, which is computed as the cumulative difference between corresponding MDCT coefficients.

The Stereo Difference block computes the Difference Level for each second as well as the Music Level. An example of the Difference Level plot is shown in Figure 4.

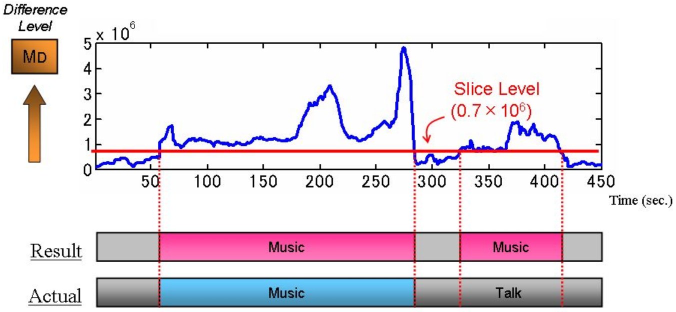


Fig. 5. Example of the Difference Level plot

Figure 5 illustrates an example of an actual broadcast program in which the segments over the slice level are determined to be the music scenes as accurately as the music level computation shown in the previous figure, but with more precise identification of the start-stop positions. As Figure 5 shows, we also have false positives since non-music segments get classified as music. The reason is that this method also detects non-music scenes that have a similar large difference between the left and right channels.

So we can say that this method finds the start/stop of a music segment with very high accuracy, but unfortunately also incorrectly flags non-musical segments as musical segments.

C. *Scene Change Support Method*

The Scene Change Detector block can detect scene change time from video feature with simple algorithms.

For example, comparing a histogram of RGB or YUV (256 bins) for each video frames can detect the positions of the shot changes. In general, the start/end time of the music in video causes large scene changes after/before interviews so a shot change is a strong cue. Since other events can also trigger shot changes, we combine this cue with the audio-based cues we described earlier to establish the start/end of musical segments.

D. *Music Segment Detection*

We introduced 2 methods of the audio based cues; Music Label Method and Stereo Difference Method. Both methods have a merit and demerit as we described in the previous, so we propose combining these 2 ways so as to compensate for their respective demerits and thus finding the music segment boundaries accurately.

Additionally, we employ Scene Change Support for refining the start/end position with frame level accuracy.

The 'Music Level', the 'Difference Level' and the 'Scene Change Time' are combined in the Music segment detection block and the system determines a music/song segment. Figure 6 shows the sequence of determination.

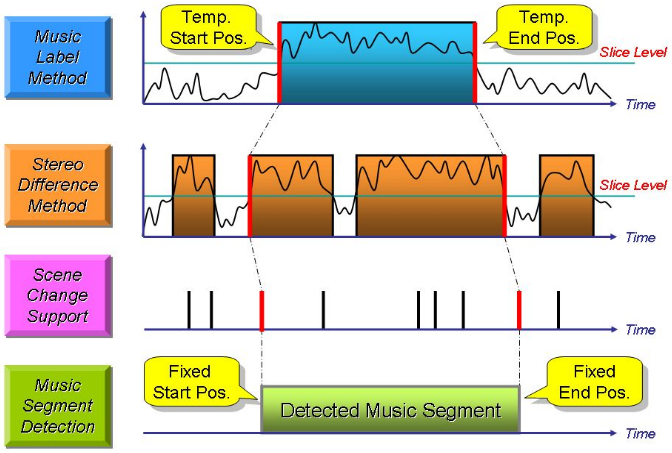


Fig. 6. Sequence of the Music Period Judgment

First, our proposed system pre-determines a music period using the 'Music Level' and searches for the start/end position of the 'Difference Level' in each neighbor. Second, the start/end positions are adjusted using the most recent 'Scene Change Time'. The combination of methods eliminates the false positives detected by the individual methods.

The music/songs periods in time (i.e. presentation time map) are stored onto the medium as a unique meta-data file. The simple feature extraction allows the meta-data generation to be done at the recording phase in real-time so user can browse the recorded content without any additional computation.

IV. PROTOTYPE DEVELOPMENT

A. *Application of Music Video Summarization*

We have developed a prototype that incorporates the proposed music segment detection system into the current HDD/DVD video recorder product without any H/W modification. Figure 5 shows a screen shot of the proposed application. The detected music segments are indicated on the Music Indicator bar with the current playing position.

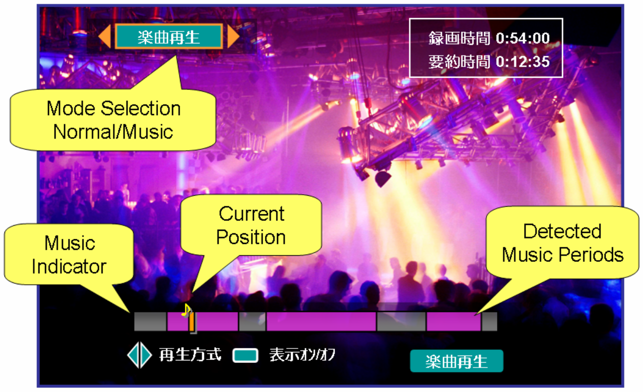
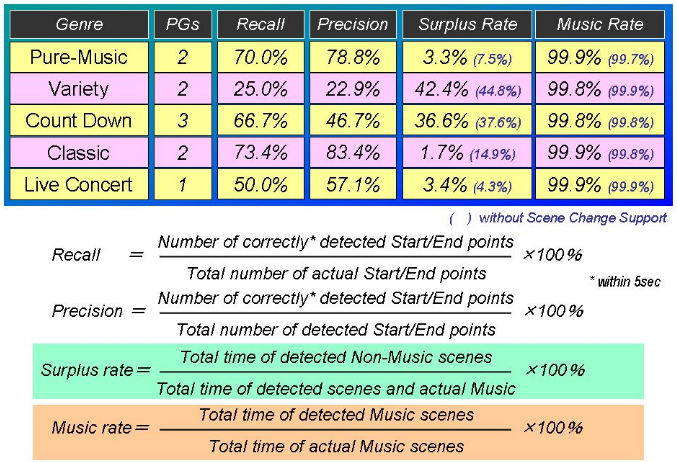


Fig. 7. An application of Music Period Browsing

When the system correctly detects the music segments, skipping to the start or end position of music interactively, would be a very useful and convenient function for the consumer. Furthermore, the system can offer skipping and playing back only the music scene automatically and can help content editors.

B. Evaluation results

We have investigated the prototype model with 10 typical Japanese music broadcast program. Figure 8 shows the ground truth.



Genre	PGs	Recall	Precision	Surplus Rate	Music Rate
Pure-Music	2	70.0%	78.8%	3.3% (7.5%)	99.9% (99.7%)
Variety	2	25.0%	22.9%	42.4% (44.8%)	99.8% (99.9%)
Count Down	3	66.7%	46.7%	36.6% (37.6%)	99.8% (99.8%)
Classic	2	73.4%	83.4%	1.7% (14.9%)	99.9% (99.8%)
Live Concert	1	50.0%	57.1%	3.4% (4.3%)	99.9% (99.9%)

( ) without Scene Change Support

$$\text{Recall} = \frac{\text{Number of correctly* detected Start/End points}}{\text{Total number of actual Start/End points}} \times 100\%$$

* within 5sec

$$\text{Precision} = \frac{\text{Number of correctly* detected Start/End points}}{\text{Total number of detected Start/End points}} \times 100\%$$

$$\text{Surplus rate} = \frac{\text{Total time of detected Non-Music scenes}}{\text{Total time of detected scenes and actual Music}} \times 100\%$$

$$\text{Music rate} = \frac{\text{Total time of detected Music scenes}}{\text{Total time of actual Music scenes}} \times 100\%$$

Fig. 8. An Evaluation Results of developed Prototype Model

We categorize the tested 10 music programs into 5 audio classes as we did in our previous work. [3]

'Pure-Music' is the traditional music program that consists of music and songs mainly and a few interviews with guests.

'Variety' consists of a couple of songs and mainly variety talk shows, otherwise short dramas, games, and so on. 'Variety' is a very popular genre in Japan.

'Count Down' is a Japanese program similar in content to shows such as 'Billboard Top 40' that mainly consist of fragmentary short music video clips.

We compute Recall and Precision, Surplus rate and Music rate. We declare a correct detection if the detected start and end positions are within 5 seconds of the ground truth. The Music rate is calculated by the total time of detected Music scenes divided by actual total time, so the rate becomes 100% when there are no misses in music detection. And the Surplus rate is the percentage of false alarms in the detected scenes, so the rate is expected to be of small value.

In Figure 8, the Music rates were almost 100% and the Surplus rates were very low, but the results for 'Variety' and 'Count Down' are worse. We think that the reason for the increased surplus rate is that the 'Variety' and 'Count Down' programs include a lot of back ground music (BGM) in non-music scenes which causes them to be falsely declared as music-segments. All in all, however, we have achieved satisfactory accuracy with our proposed method. Note that the number in parentheses indicates the result without the scene change support. Our results clearly indicate that the scene change support improves the Surplus rate.

V. Conclusion

We have developed a system that can detect music/song segments with high accuracy. Analyzing music scenes by audio features with the support of video scene change detection achieves accurate detection of music segments. Our approach employs a combination of very low-complexity audio and video feature extraction. Our enhancements will therefore be easy to incorporate into the target platform. Extension of our framework to other genres is an avenue for further improvement.

Acknowledgment

The authors greatly appreciate the valuable contribution of Dr. Regunathan Radhakrishnan and Mr.Isa for their consistent encouragement as well as the encouragement from our colleagues at Kyoto Works, Mitsubishi Electric Corporation.

References

[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.

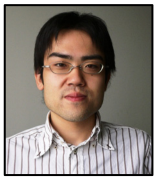
[2] I. Otsuka, K. Nakane, A. Divakaran, K. Hatanaka and M. Ogawa, "A Highlight Scene Detection and Video Summarization System using Audio Feature for a Personal Video Recorder", *IEEE Transaction on Consumer Electronics*, vol.51-1, Feb.2005, pp. 112-116

[3] I. Otsuka, R. Radhakrishnan, M. Siracusa, A. Divakaran and H. Mishima, "An Enhanced Video Summarization System using Audio Features for a Personal Video Recorder", *IEEE Transaction on Consumer Electronics*, vol.52-1, Feb.2006, pp. 168-172

Isao Otsuka received his B.E. degree in precision mechanical engineering from Meiji University, Japan in 1989.

He joined Mitsubishi Electric Corp. in 1989 and has been engaged in the research and development of speaker system and sound field analysis for home theater and car audio applications.

Recently, he has been engaged in the development of system for storage devices such as DVD & HDD recorders/players. His interests span applications for digital storage devices as well as relevant core technologies such as audio-visual content analysis. He has recently published several conference papers on video summarization and audio classification. He is a member of the Institute of Image Information and Television Engineers.



Hidetsugu Suginohara received his B.E. degree and M.E. degree in electronics engineering from Osaka University, Japan in 2003 and 2005, respectively.

He joined Mitsubishi Electric Corp. in 2005 and has been engaged in the research and development of the video summarization system for storage devices, such as DVD/HDD recorders and Blu-ray Disc recorders, at the Advanced Technology R&D Center.

Yoshiaki Kusunoki received his B.E. degree and M.E. degree in mechanical and system engineering from Kyoto Institute of Technology University, Japan in 1991 and 1993, respectively.

He joined Mitsubishi Electric Corp. in 1993 and has been engaged in research and development of MPEG-2 codec technologies, DVD authoring systems and video on demand systems. At present, he is engaged in the development of storage equipment , such as DVD/HDD recorders and Blu-ray Disc recorders, at Advanced Technology Center. Mitsubishi Electric Corp.

Ajay Divakaran (SM'00) received the B.E. (with Hons.) degree in Electronics and Communication Engineering from the University of Jodhpur, Jodhpur, India, in 1985, and the M.S. and Ph.D. degrees from Rensselaer Polytechnic Institute, Troy, NY in 1988 and 1993respectively. He was an Assistant Professor with the Department of Electronics and Communications Engineering, University of Jodhpur, India, in 1985-86. He was a Research Associate at the Department of Electrical Communication Engineering, Indian Institute of Science, in Bangalore, India in 1994-95. He was a Scientist with Iterated Systems Inc., Atlanta, GA from 1995 to 1998. He joined MERL in 1998 and is now a Senior Team Leader - Senior Principal Member of Technical Staff. He has been an active contributor to the MPEG-7 video standard. His current research interests include video and audio analysis, summarization, indexing and compression, and related applications. He has published several journal and conference papers, as well as six invited book chapters on video indexing and summarization. He has co-supervised four doctoral theses. He currently serves on program committees of key conferences in the area of multimedia content analysis. He has also coauthored a book titled “A Unified Framework for Video Summarization, Browsing and Retrieval” (Elsevier Academic Press).