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

An Enhanced Video Summarization System Using Audio Features for a Personal Video Recorder

Isao Otsuka, Regunathan Radhakrishnan, Michael Siracusa, Ajay Divakaran, and Hidetoshi Mishima

TR2006-024 February 2006

Abstract

We extend our Sports Video Browsing framework for Personal Video Recorders, such as DVD Recorders, Blu-ray Disc Recorders and/or Hard Disc Recorders, to other genres. We reduce the computational complexity by reducing the number of audio classes to a small but useful set that is useful for both sports video and music video, as well as by reducing the complexity of the Gaussian Mixture Models. Our extension to music video content consists of detecting music/song periods by compensating for false alarms. Our results indicate that our enhanced audio-only summarization maintains the sports video performance and works well with music video content. We can therefore integrate the enhancement into our product while in fact reducing the computational complexity.

IEEE Transactions on Consumer Electronics

An Enhanced Video Summarization System using Audio Features for a Personal Video Recorder

Isao Otsuka, Regunathan Radhakrishnan, Michael Siracusa, Ajay Divakaran, and Hidetoshi Mishima

Abstract — *We extend our Sports Video Browsing framework for Personal Video Recorders, such as DVD Recorders, Blu-ray Disc Recorders and/or Hard Disc Recorders, to other genres. We reduce the computational complexity by reducing the number of audio classes to a small but useful set that is useful for both sports video and music video, as well as by reducing the complexity of the Gaussian Mixture Models. Our extension to music video content consists of detecting music/song periods by compensating for false alarms. Our results indicate that our enhanced audio-only summarization maintains the sports video performance and works well with music video content. We can therefore integrate the enhancement into our product while in fact reducing the computational complexity.* [1]

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

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]. Current PVRs can store up to 200 hours of high quality video content, and the storage capacity is expected to grow further.

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], [3], 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.

In this paper we address two challenges. First, we reduce the complexity of the audio classification by reducing the number and size of the Gaussian Mixture Models for a realization of dual channel extraction system. Second, we extend the summarization to music videos by using audio segmentation based on audio-classification and low-level audio feature analysis.

[1] Isao Otsuka, Hidetoshi Mishima are with the Advanced Technology R&D Center, Mitsubishi Electric Corporation, Kyoto, Japan (e-mail: Otsuka.Isao@cw.MitsubishiElectric.co.jp, Mishima.Hidetoshi@dn.MitsubishiElectric.co.jp)

Regunathan Radhakrishnan, Michael Siracusa, and Ajay Divakaran are with Mitsubishi Electric Research Laboratories, Cambridge, USA (e-mail: regu, siracusa, ajayd@merl.com).

II. PROPOSED SYSTEM FOR SPORTS VIDEO SUMMARIZATION

A. Application Framework

When we focus on the sports video program, interesting events lead to human reaction consisting of a mixture of cheering and the commentator's excited speech. Thus automatic detection of excited speech parts in the content provides quick access to the highlight scenes and we propose a video browsing system based on this information.

A basic concept of the proposed sports video browsing system is shown in Figure 1.

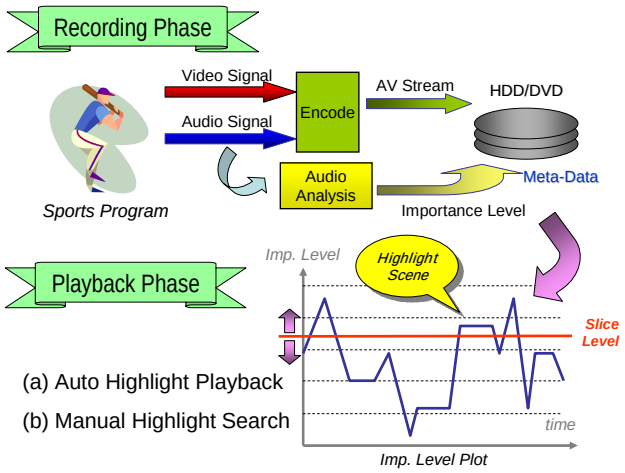


Fig. 1. Basic Concept of the Proposed Sports Video Summarization.

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 analyses audio features and classifies the audio class in real time by using a digital signal processor (DSP). Then the system calculates an Importance Level for each second using the extracted features, such as an appearance order and continuation of the specific audio class, sound level, and so on. Our proposed system uses only audio features, which means the calculation complexity is lower than that of a system that uses video features. The calculated Importance level data is stored onto 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 importance level which is gotten from Meta-data can be plotted with a slice

level as shown in Fig.1. The segments over the slice level are determined to be the highlights. The slice level is adjustable by the user for changing the playback time. Thus we propose the following functions; Skipping to the start position of the highlight scene manually is the function of Highlight Search, and skipping and playing back only the highlights automatically is Auto Highlight Playback.

The audience response to interesting events in different sports is same irrespective of the type of interesting event, e.g. goal scoring in soccer, grand slam in baseball, and so on. Therefore, a single audio analysis framework which measures the reaction can work across a wide range of sports, as well as languages and commentators, and broadcasting companies.

B. System Configuration

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 2.

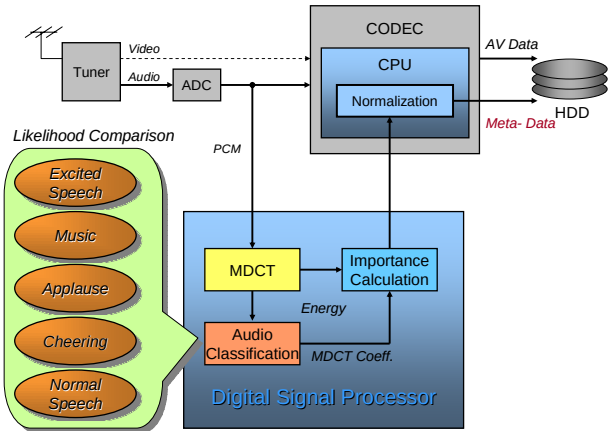


Fig. 2. Simplified Block Diagram for the Recording Phase.

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 content. So the system classifies the input audio by comparing the likelihoods of the audio classes as shown in Figure 3. 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.

The Importance calculation block computes the percentage of excited speech in a segment for sports video, and calculates the audio energy of the segment, then multiplies these two to get an Importance Level for a video shot. This method also can check for segment boundaries in the case of music video.

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.

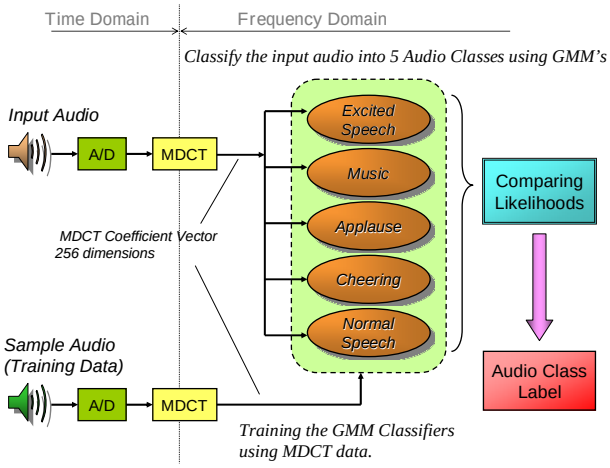


Fig. 3. Audio Class Recognition (GMM: Gaussian Mixture Model).

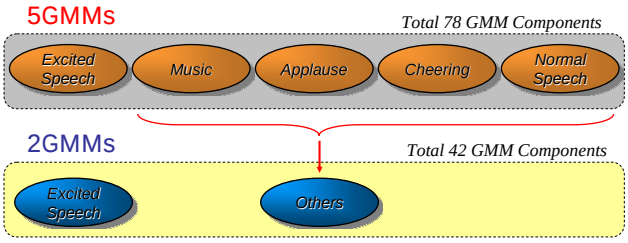
C. Increasing Calculation Speed

In the development of the sports video summarization, some issues remain. In Japanese market, a double tuner model that can record the different programs simultaneously has become popular. So, dual channel extraction has to be supported for such models.

To realize dual channel extraction, increasing the calculation speed is essential. We mentioned that our proposed system calculates the importance level in real time which means the system must calculate MDCT coefficients and extract an importance level within an audio frame that has a duration of 32msec. Currently our system takes 19.6msec in all for a channel, by using a 300MHz fixed point DSP. Since the MDCT coefficient calculation that has 2.8msec itself is almost optimized, now we focus on reducing the time taken to compute the importance level. The condition to extract two channels within the audio frame, the calculation time for a channel should be under 32/2=16msec. in all.

We are employing five audio models (5GMMs) for audio classification as we mentioned in this paper. The importance level calculation part takes 16.8msec so far and it consists of 'MDCT Coefficient Calculation', 'Likelihood Calculation', 'Importance Level Calculation' and some overheads.

To solve that problem, we employed two audio models (2GMMs) by combining 4 of the previous 5 audio classes 'Applause', 'Cheering', 'Music' and 'Normal Speech' to 'Others' as shown in Figure 4.



	5GMMs	2GMMs
MDCT Coefficient Calculation	3.5 msec.	3.5 msec.
Likelihood Calculation for each GMMs	13 msec.	7 msec.
Importance Level Calculation	> 0.1 msec.	> 0.1 msec.
Other Overheads	0.3 msec.	0.3 msec.
Total	16.8 msec.	10.9 msec.

Calculation Time (@300MHz Fixed Point DSP)

Fig. 4. GMM Components Reduction.

The calculation complexity depends on the total number of Gaussian components. 5GMMs have 78 Gaussian components totally but 2GMMs have only 43 Gaussian components which save us 35% of the computations. We can thus support the dual extraction.

Figure 5 is a comparison of the confusion matrix between 5GMMs and 2GMMs. The accuracy in fact goes up when we use 2GMMs. It seems surprising, however, what we have in fact done is changing the approach from a five way classification to a binary classification. The binary classification is an easier task which can therefore be achieved with both higher accuracy and lower complexity. Note that the five way classification gives us much more information about the audio scene than the binary classification approach.

5GMMs
Order = Applause, Cheering, Music, Speech, Excited Speech
# Components = 7 , 18 , 11, 28 , 14 (Total: 78)
confusionMatrix = 0.7018 *0.0515* *0.1245* *0.0630* *0.0592*

$$\begin{pmatrix} 0.0305 & 0.8057 & 0.0495 & 0.0229 & 0.0914 \\ 0.1825 & 0.0298 & 0.7181 & 0.0478 & 0.0219 \\ 0.0426 & 0.0578 & 0.0456 & 0.7705 & 0.0836 \\ 0.0333 & 0.1067 & 0.0178 & 0.0978 & 0.7444 \end{pmatrix}$$

Detection Accuracy = 74.8%

2GMMs
Order = Excited Speech, Other
# Components = 14, 28 (Total: 42)
confusionMatrix = $\begin{bmatrix} 0.8044 & 0.1956 \\ 0.0944 & 0.9056 \end{bmatrix}$
Detection Accuracy = 85.5%

Fig. 5. Comparison of Confusion Matrix between 5GMMs and 2GMMs.

Figure 6 shows a comparison of importance level plot between 5GMMs and 2GMMs for the 2002 Soccer World Cup Final. 6 highlight scenes were extracted from each set of models and we can see that they are almost the same.

Thus, reducing the number of GMMs can increase calculation speed while getting essentially the same result.

Fig. 6. Comparison of Importance Level Plot between 5GMMs and 2GMMs.

III. Proposed System for Music Video Summarization

A. Application Framework

In this section, we propose to extend the browsing application based on the importance level calculation of a specified audio class to other genres. In this paper we propose an adoption to music program that can detect music or song periods from the entirely recorded music program. Figure 7 shows a screen shot of the music video browsing application.

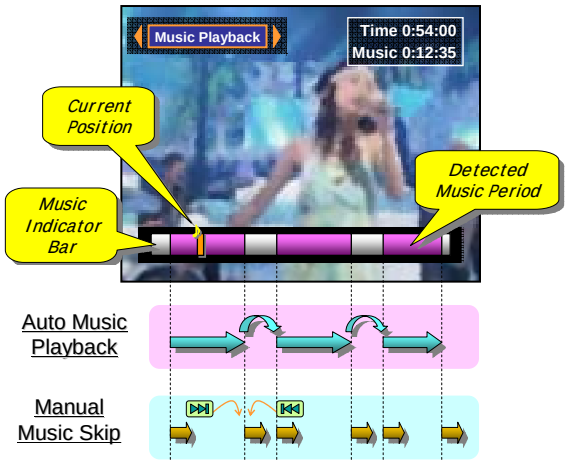


Fig. 7. An Application of Music Video Summarization.

Just like the sports video summarization, the system detects the music boundaries from the audio signal at the recording phase, and then stores the information of detected music positions to HDD or DVD as Meta-data. The detected music periods are indicated as shown in Figure 7 on the Music Indicator Bar with the current playing position.

When the system correctly detects the music periods, skipping to the start or end position of music interactively would be a very useful and convenient function. And also the system can offer skipping and playing back only the music

scenes automatically.

B. *Method of Music Detection*

Figure 8 shows an importance level plot that was extracted from a typical Japanese music program with Music class. There are 5 music segments and 3 talk shows, and there are some commercial messages in the recorded program.

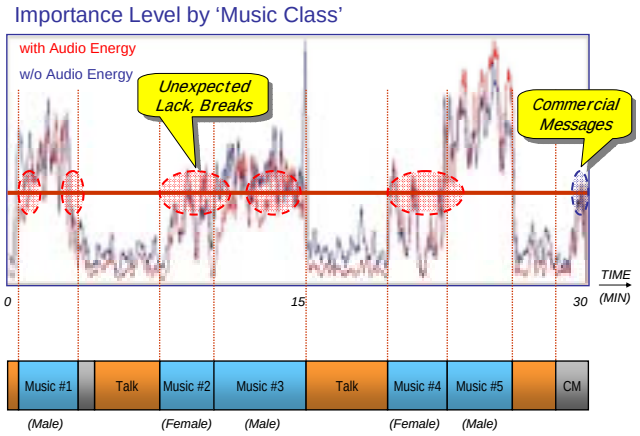


Fig. 8. An Example of Importance Level Plot from of typical Japanese Music Program.

Importance level calculation is good for searching sporadic highlight moments in sports video. But for music video, we are interested in identifying whole periods of music/song. Even if the system can extract an importance level of music periods by same manner of sports highlight detection, it will have some unexpected breaks and unwanted segments such as commercial messages.

Therefore we propose to employ the median filters as shown in Figure 9, for eliminating undesired segments.

[Filter #1] Reject the Detected Commercial Message periods.

[Filter #2] Eliminate any breaks (< 3sec.) in Music periods.

[Filter #3] Fix a Music period which has over 90 sec.

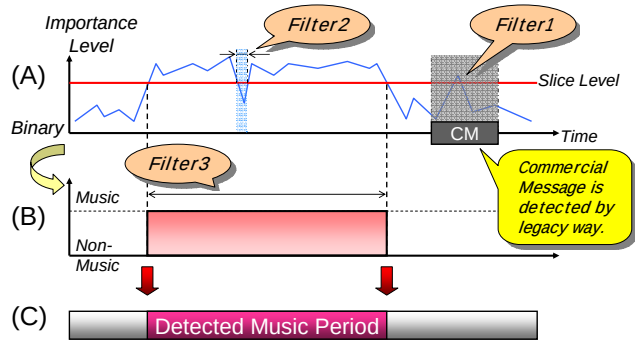


Fig. 9. Median Filters for compensation.

Figure 9 (A) is showing an example of Importance level plot, the horizontal axis is the playback time and the vertical axis is the value of the Importance level based on Music class with slice level.

First, the filter #1 rejects the commercial message periods which are detected using a legacy system. For example, a commercial message period can be identified by detecting the moment of the audio mode changing from STEREO to MONO, and so on.

Second, median filtering in music segments to prevent spurious non-music sections. Filter #2 eliminates any breaks less than 3 seconds as shown in Figure 9 (B).

Third, median filtering in non-music segments to prevent spurious music sections. We defined that a compensated period which has over 90 seconds should be fixed as 'Music'. So Filter #3 rejects music sections lasting less than 90 seconds. Then the music periods are identified as shown in Figure 9 (C).

C. *Prototype Model Development*

We have developed a prototype model that can detect music periods automatically in real time and playback desired scenes by 'Auto Music Playback' and 'Manual Music Skip' functions that were described in Figure 7.

We have tested the prototype model with 10 typical Japanese music TV programs initially. Figure 10 shows the ground truth.

Genre	PGs	Music	Recall	Precision	Surplus rate	Music rate
Pure-Music	2	10	65.0%	65.0%	2.5%	87.2%
Variety	2	6	91.7%	91.7%	8.3%	98.7%
Count Down	3	21	83.3%	83.3%	3.8%	91.8%
Live Concert	1	8	62.5%	71.4%	1.4%	91.8%
Classic	2	8	37.5%	42.9%	4.6%	77.4%

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

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

* within 10sec

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

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

Fig. 10. An Evaluation Results of developed Prototype Model.

We categorize the tested 10 music TV programs to 5 genres.

'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 very popular genre in Japan.

'Count Down' is a program like a 'Billboard Top 40' that consists of fragmentary short music video clips mainly.

We checked Recall and Precision, Surplus rate and Music rate. In Recall and Precision, we counted as OK when the detected start positions and end positions are within 10sec. against the actual music periods.

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 Surplus rate is percentage of false alarms in the detected scenes, so the rate is expected to be of small value.

In Figure 10, the Music rates were almost over 90% and the Surplus rates were very small. 'Classical' music has silent parts sometimes so the Recall, Precision, and Music rate worsen. Furthermore, 'Variety' includes a lot of back ground music (BGM) in non-music scenes so the Surplus rate increases. All in all, however, our proposed method provides satisfactory accuracy.

IV. CONCLUSION

We have developed a system that extracts highlight scenes automatically by using audio features and plays back desired highlight scenes. And we have developed a dual channel extraction system for double TV tuner model by reducing the number of Gaussian components. We have retained the audio –only approach, and greatly reduced the complexity of the audio calculation. Additionally, we developed a music detection system that can play back music and song scenes with high accuracy. Our enhancement will therefore be easy to incorporate into the target platform. Finally, extension of our framework to other genres offers avenues for further improvement.

ACKNOWLEDGMENT

The authors greatly appreciate the valuable contribution of Mr. Masami Isa, and Mr. Shigenori Suginohara to studying and prototyping the system described in this paper and Dr. Masaharu Ogawa for his consistent encouragement as well as 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] 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.

[3] 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", ICCE, 2005, Las Vegas.

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 and man-machine interface 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.

Regunathan Radhakrishnan received the B.E (Hons) in EE and M.Sc (Hons) in Chemistry from Birla Institute of Technology and Science (BITS), Pilani, India in 1999. He worked as DSP Engineer in Multimedia Codecs Group at SASKEN Communication Technologies Ltd, Bangalore, India in 1999-2000. He received the M.S. and Ph.D in EE from Polytechnic University, Brooklyn, NY in 2002 and 2004 respectively. He was a research fellow in the ECE department and also an intern at Mitsubishi Electric Research Labs, Cambridge, MA during his graduate studies. He joined Mitsubishi Electric Research Laboratories (MERL) in 2005 as a Visiting Researcher. His Current research interests include audio classification, video summarization, fault detection and diagnosis, digital watermarking & content security and data mining. He has published several conference papers, as well as 5 journal papers and 3 book chapters on multimedia content analysis and security. He has also coauthored a book titled "A Unified Framework for Video Summarization, Browsing and Retrieval" (Elsevier Academic Press).

Michael Siracusa received a B.S. in Computer Engineering from the State University of New York, Stony Brook, in 2002. He is currently a graduate student working towards his Ph.D. at the Massachusetts Institute of Technology. His research interests include machine perception and learning, and higher level processing of multi-modal data such as audio-visual event detection and general scene analysis. He worked as a summer research intern at Mitsubishi Electric Research Laboratories in 2005.

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 ofElectronics and Communications Engineering, University of Jodhpur, India, in 1985-86. He was a Research Associate at the Department ofElectrical 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 andcompression, 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).

Hidetoshi Mishima received the B.E. degree in 1986 from Osaka University, Japan.

He joined Mitsubishi Electric Corp. in 1986, where he has been engaged in the development of consumer VCRs, MPEG-2 encoders, DVD authoring systems and video on demand systems. At present, he is engaged in the development of storage equipment, such as DVD recorders and Blu-ray recorders, and home servers by using home network technology, at Advanced Technology R&D Center, Mitsubishi Electric Corp..

He is a member of the Institute of Electronics Information and Communication Engineers, the Institute of Image Information and Television Engineers, and the Institute of Image Electronics Engineers of Japan. He received Imaging and Visual Computing Technology Award in 2005 from the Institute of Image Electronics Engineers of Japan.