Multimedia Navigation Technology for Handling Multimedia Flexibly

Masashi Morimoto†*, Shozo Azuma, Hiroshi Konishi, and Satoshi Shimada*

Abstract

This article presents multimedia navigation technologies for portal services that enable users to easily arrange, search, comprehend, and publish data, including videos, as desired.

1. Need for multimedia navigation

The use and distribution of multimedia contents such as images and videos on the network are rapidly increasing due to the development of broadband networks, decreasing size of capturing/recording systems such as digital videos/cameras and solid-state memory recorders, lower costs of home digital equipment such as personal computers (PCs) for storing and viewing content, and the proliferation of camera-equipped phones and mobile music players. However, it is still time-consuming for users to categorize and arrange multimedia information according to its content so that they can see what is where, process and edit multimedia contents to share them with others, or find the data that they want amid a huge volume of multimedia contents. Therefore, by creating appropriate multimedia navigation technologies, we can create new multimedia portal services that support not only information searching on the network, but also the activities of individuals and communities.

2. Outline of multimedia navigation technologies

Video is widely expected to rapidly become the predominant medium used in portal services in the future. Typical actions involving video are distribution and searching. However, users find it especially difficult to understand the contents of video files quickly and handle video because video files can take a long time to play and can contain a lot of information.

NTT Cyber Solutions Laboratories has developed video indexing technology and metadata generation technology, which enable users to comprehend the content to a reasonable extent because it is indexed by content analysis and extracts [1], [2]. Building on basic technologies such as these, we are now developing multimedia navigation technologies that enable users to easily manipulate data as desired. The usage style is likely to change with the time and situation. For example, users will want the huge volume of video files stored on their PC to be indexed and arranged automatically to enable them to grasp their contents at a glance; they may capture extensive footage of a particular event using a camera-equipped phone and then edit it and publish only a few good scenes; and they may want to talk about interesting videos by referring to specific scenes or discuss and recommend actors, places, and things in those scenes. Multimedia navigation technologies can provide these functions as aspects of new portal services (**Fig. 1**).

3. Examples of multimedia navigation technologies

This section presents three examples of multimedia navigation technologies being developed by NTT: Video Pot, MobileMovieClub, and SceneNAVI.

3.1 Categorize and arrange videos without any special expertise—"Video Pot"

More individuals are using videos than ever before.

† NTT Cyber Solutions Laboratories
Yokosuka-shi, 239-0847 Japan
E-mail: morimoto.masashi@lab.ntt.co.jp

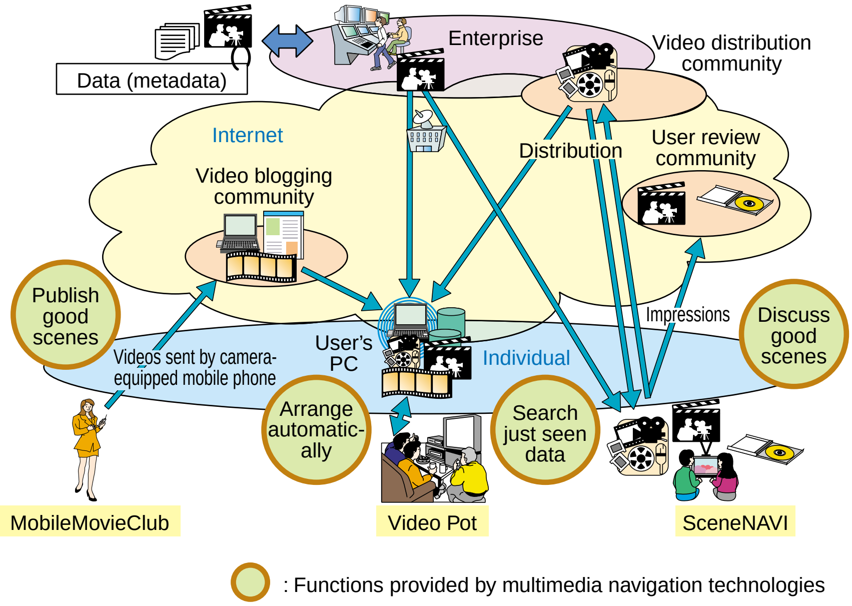


Fig. 1. Multimedia navigation technologies, functions, and service concepts.

Video Pot lets them store videos recorded from television, scenes they have shot on camera themselves, or data downloaded from the Internet, etc., on their PCs as video files of various applications. Users can find the video they want from among these files based on specific information such as:

(a) a scene they remember or

(b) a keyword directly related to the video (such as the title, a performer, or the release date of the video)

If the users do not have any specific information, their search can be based on nonspecific information such as:

(c) information that might be related to the video

In case (a), users visually search for the video based on its scene characteristics; for (b), they use keywords representing the video's characteristics; and for (c), they must conduct a visual search based on an outline of the video or use other users' comments. To search for a video using any of these three methods ((a)-(c)), users must enter keywords or identify scene characteristics in advance. Therefore, the burden on users increases with the number of video files.

To solve such problems, several papers have addressed the issues related to video indexing based on multiple audio and video features [2], and relevance feedback technologies and interfaces for effective video searches have been recent topics in video retrieval [3]-[6]. However, compared with popular desktop search tools, they are not so simple and have a high operation cost. Desktop search tools using metadata based on the history of users' actions have been proposed [7], [8]. These studies focused on using the history of accessed Web pages and the history of a file's location in a file system to search for files. However, that information is not directly related to the content of multimedia files.

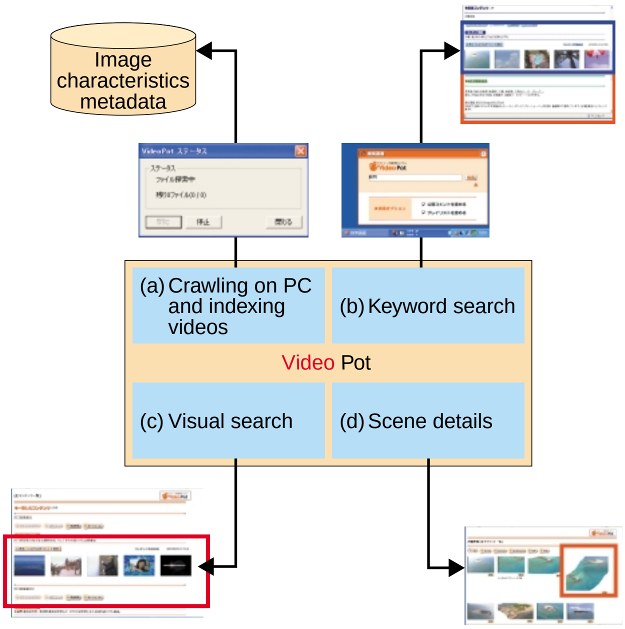


Image characteristics metadata
(a) Crawling on PC and indexing videos
(b) Keyword search
Video Pot
(c) Visual search
(d) Scene details

Fig. 2. System outline of Video Pot.

We are developing Video Pot, a desktop video search system with automatic video categorization and arrangement technology (**Fig. 2**) [12]. Video Pot crawls through the hard drive on a user's PC, identifies all video files, and automatically prepares indexes of scenes that characterize them by extracting video features [9]-[11]. It also registers the metadata information related to each video, such as file name, file properties, and information in electronic program guides (TV guides), as metadata (keywords). After these processes, Video Pot automatically arranges and manages all video files on the PC ((a) in Fig. 2).

Video Pot has two search functions: searching by keyword and visual searching. These let users more easily search for the video they want from among all the available video files. The keyword search lets users comprehensively search for video files on the Internet at the same time as searching for ones on their own PCs ((b) in Fig. 2). Thus, search results from the Internet can complement desktop search results. Moreover, if users cannot find the target video file, they will be given candidate keywords for another search attempt. The visual search displays multiple frame images representing each video file, which enables the user to judge whether a particular video is the one he/she is looking for ((c) in Fig. 2). For each candidate video file, Video Pot shows various types of index information (frame image characteristics of the video in terms of scene-changes, captions, camera operations, sound, music, etc.) so that users can discover more details about the video content without having to play it ((d) in Fig. 2).

3.2 Easily publish selected scenes of videos —"MobileMovieClub"

Recently, blogging has become a quickly spreading passion among Internet users. With the explosion of camera-equipped mobile phones, blog (Weblog) systems have expanded to encompass mobile blogging (moblog) and video blogging (vlog) systems [13]. However, even current mobile blogging systems, which make it easy for users to post video by using the phone's built-in camera and e-mail, provide poor usability in terms of video browsing and editing. For example, they link to videos but show only the first frame of each one. Therefore, users cannot see the content without playing the entire video, which is rather inconvenient for browsing. Moreover, while bloggers are happy to write about their feelings and opinions, they may feel uncomfortable making their videos available to the public, but editing a video to make it suitable for public viewing is bothersome. Thus, there are several factors that make individuals reluctant to release their videos to the public.

To resolve this issue, we are developing MobileMovieClub (MMC), a video community blog system mainly for mobile phone users that uses personal media profiling [14]. In addition to the ordinary video blog function, MMC incorporates video indexing technology that generates thumbnail or panoramic images that represent the video. MMC enables users to select good scenes from the videos and write an article (blog entry) in chronological order by using multiple thumbnail images or write an article that describes an extended space by using panoramic images. In this way, users can overview video contents and write expressive articles without playing back the videos. The video indexing technology also provides a simple editing function that can skip parts of a video by detecting the video segments and designating the ones that should be played. Furthermore, while conventional video blogs can be seen by anyone, MMC lets users limit access to approved users only. Since the community-control technology enables users to share posted videos among friends, edit articles by themselves, and limit access to their articles, it helps activate video blog communities through the stream of information published, as shown in **Fig. 3**.

These characteristics are very effective, especially when used in conjunction with mobile phones with cameras. MMC not only allows users to post videos shot by camera-equipped phones to the MMC server easily as e-mail attachments, but also lets them edit video segments and blog articles via the browser running on the mobile phone. Therefore, MMC enables users to publish in an effective way by editing and posting the videos they shot on the spot or articles they wrote. Also, when users browse video blogs from the phone's browser, they can overview the images without having to download the entire videos, so they can efficiently find the article they want, as shown in **Fig. 4**.

3.3 Discuss interesting scenes—"SceneNAVI"

We often want to discuss the content of videos with other people. This can be done synchronously by using a realtime chat system to communicate with others while you are watching the video or asynchronously by posting a message to a bulletin board system (BBS). Synchronous communication seems more natural to users, but they often drift off topic and are unable to discuss one topic deeply. On the other hand, asynchronous communication lets users discuss a certain topic in detail, but they miss the feedback of normal conversation and readers may get confused about which scene is being discussed.

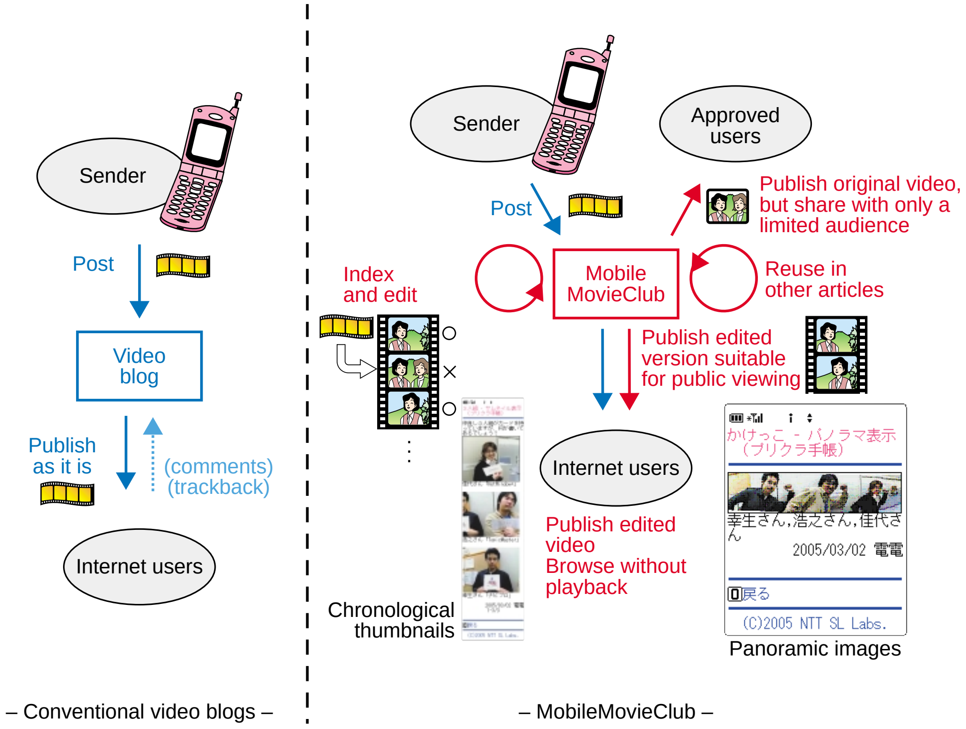


Fig. 3. MobileMovieClub overcomes problems of conventional video blog.

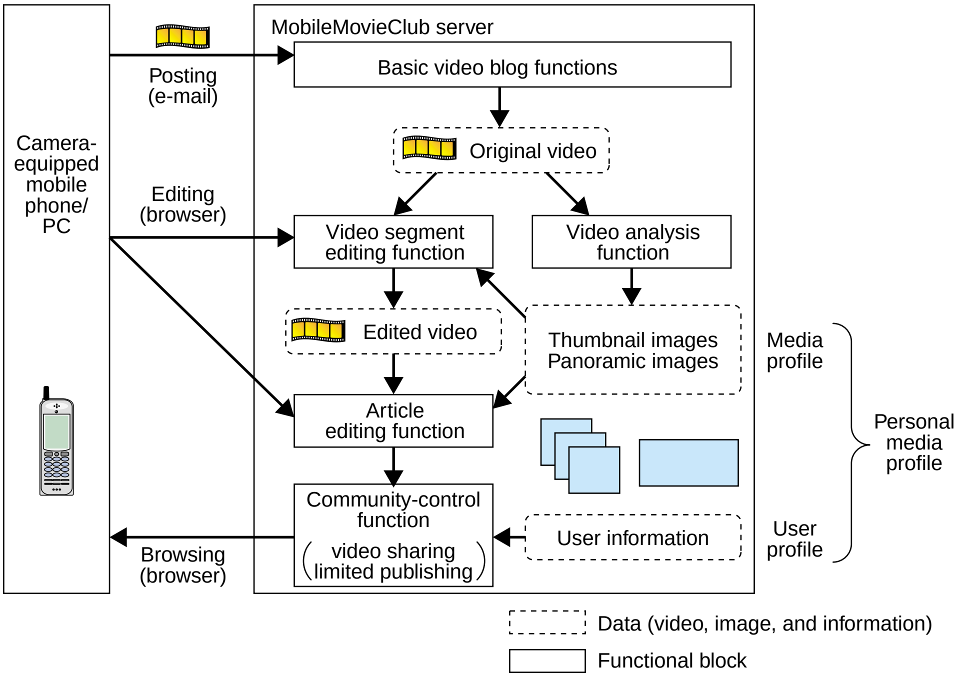
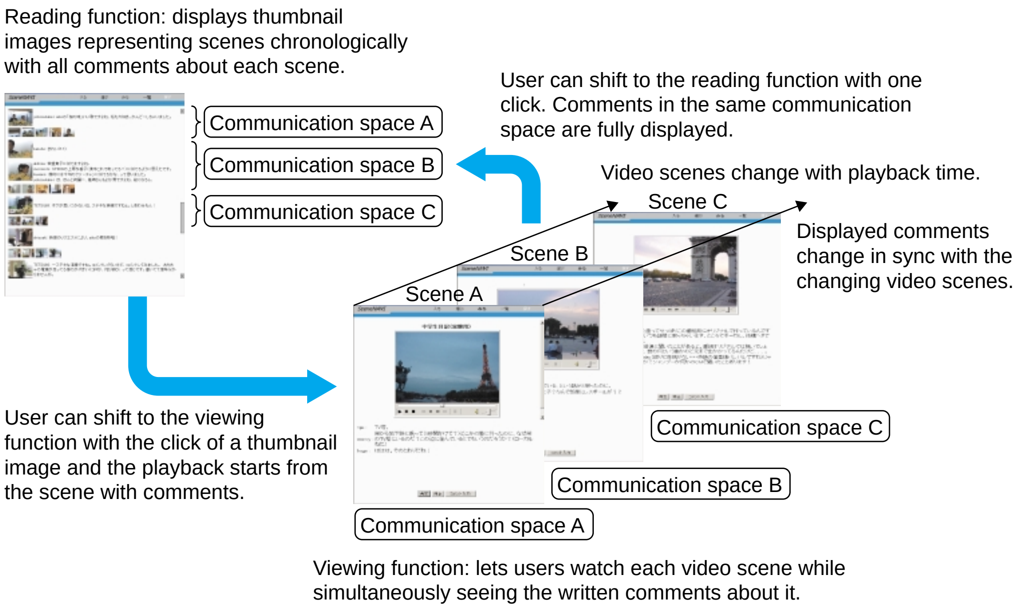


Fig. 4. System outline of MobileMovieClub.

To solve these problems, we are developing SceneNAVI, a communication system that uses video-linked communication technology. It combines video viewing and communication functions for registering or browsing comments about a video scene, which is segmented to match the written comments [15], [16].


Reading function: displays thumbnail images representing scenes chronologically with all comments about each scene.

User can shift to the reading function with one click. Comments in the same communication space are fully displayed.

Video scenes change with playback time.

Displayed comments change in sync with the changing video scenes.

User can shift to the viewing function with the click of a thumbnail image and the playback starts from the scene with comments.

Viewing function: lets users watch each video scene while simultaneously seeing the written comments about it.

Fig. 5. SceneNAVI: viewing and reading functions.

3.3.1 Basic functions of SceneNAVI

SceneNAVI creates a communication space for each video scene defined beforehand. The communication space is established by either the viewing function or the reading function, as shown in **Fig. 5**.

The viewing function allows users to communicate mainly by watching and listening to videos. It creates a single communication space by synchronizing the video scenes being watched by multiple users. For example, it shows communication space A where viewers can read or write comments about scene A while scene A is being played back. In this way, SceneNAVI presents one by one communication spaces synchronized to the part that is currently being watched. This gives the users a realistic feeling as if they were communicating with each other simultaneously.

The reading function allows users to communicate with others via comments. The communication space of a scene contains comments made about that scene together with thumbnail images with titles representing the scene. This function lists such communication spaces. In this way, users can understand the relationship between the scene and the comment in order to better understand the overall situation or to have a deep discussion in the same manner as on a BBS. By combining these two functions, users can read topics that interest them while viewing the video or replay scenes of interest while reading comments about them.

3.3.2 Navigation function

Communication through videos offers users an opportunity to share topics with others easily, which helps active communication among users with similar interests. Therefore, it is important to establish a system to activate the community by showing users the scenes from videos that are interesting to them.

We are therefore developing community-oriented navigation technology to assist users in encountering scenes of interest to them. It automatically generates scene profiles or metadata of video scenes from comments registered by users and user profiles from user behavior histories. Both profiles will be appropriately and dynamically presented to a user according to his/her use. More precisely, this technology automatically generates access metrics such as scene access scores or topical words according to the comments registered by the user or the user's history in accessing SceneNAVI. It selects for presentation the items suitable for the user and his/her usage status of the provided functions such as a function for selecting contents from a list, the viewing function, and the reading function.

4. Future work

This article presented an overview of multimedia navigation technologies. It introduced three representative technologies: automatic video categorization and arrangement, personal media profiling, and video-linked communication with video indexing. Besides these technologies, content-search technology is also important to enable users to search what they just saw. We will continue to develop advanced multimedia search technologies that allow users to search for video scenes based on content, such as "Multimedia Meister", which was described in the previous article in this Special Feature [17]. Moreover, regarding the basic function of video indexing, we will further develop integrated multimedia indexing technologies that comprehensively use multiple information sources to make indexes that better express the meaning and content of videos.

References

[1] M. Tsunakara, R. Kataoka, and M. Morimoto, "Framework for Supporting Metadata Services," NTT Technical Review, Vol. 1, No. 3, pp. 57-61, 2003.

[2] H. Kuwano, Y. Kon'ya, T. Yamada, and K. Kawazoe, "SceneCabinet/Live!: Realtime Generation of Semantic Metadata Combining Media Analysis and Speech Interface Technologies," NTT Technical Review, Vol. 3, No. 8, pp. 40-46, 2005.

[3] TREC Video Retrieval Evaluation http://www-nlpir.nist.gov/projects/trecvid/

[4] M. Christel and N. Moraveji, "Finding the right shots: Assessing usability and performance of a digital video library interface," Proc. of ACM Multimedia 2004, pp. 732-739, 2004.

[5] A. Girgenson, J. Adcock, M. Cooper, and L. Wilcox, "Interactive search in large video collections," Proc. of ACM Conf. on Human Factors in Computing Systems (CHI 2005), pp. 1395-1398, April, 2005.

[6] Y. Kinoshita, N. Nitta, and N. Babaguchi, "Interactive clustering of video segments for media structuring," Proc. of IEEE Int. Conf. on Multimedia and Expo (ICME 2005), pp. 630-633, 2005.

[7] P. Chirita, R. Gavriloaie, S. Ghita, W. Nejdl, and R. Paiu, "Activity based metadata for semantic desktop search," Proc. of 2nd European Semantic Web Conf. (ESWC2005), pp. 439-454, 2005.

[8] T. Morita, T. Hidaka, T. Kura, K. Ooura, and Y. Kato, "Desktop search system based on the action-oriented algorism," (*sic*) Proc. of 6th Asia-Pacific Symposium on Information and Telecommunication Technologies (APSITT 2005), pp. 204-207, 2005.

[9] K. Minami, A. Akutsu, H. Hamada, and Y. Tonomura, "Video Handling with Music and Speech Detection," Proc. of IEEE Multimedia 98, Vol. 5, No. 5, pp. 17-25, 1998.

[10] H. Kuwano, Y. Taniguchi, H. Arai, M. Mori, S. Kurakake, and H. Kojima, "Telop on Demand: Video Structuring and Retrieval based on Text Recognition," Proc. of IEEE Int. Conf. on Multimedia and Expo 2000, pp. 759-762, 2000.

[11] Y. Taniguchi, A. Akutsu, and Y. Tonomura, "PanoramaExcerpts: Extracting and Packing Panoramas for Video Browsing," Proc. of ACM Multimedia 97, pp. 427-436, 1997.

[12] H. Nagata, D. Mikami, S. Azuma, and M. Morimoto, "VideoPot: Indexing-based Desktop Video Search System," Proc. of IEEE Int. Conf. on Multimedia and Expo 2006, pp. 373-376, 2006.

[13] C. Parker and S. Pfeiffer, "Video Blogging: Content to the Max," IEEE MultiMedia, Vol. 12, No. 2, pp. 4-8, 2005.

[14] H. Konishi, Y. Torii, and M. Morimoto, "Mobile Video Blog System for Mobile Camera Phones," Proc. of the 2006 IEICE General Conference, B-15-1, pp. 563, 2006 (in Japanese).

[15] K. Yamada, K. Miyakawa, and M. Morimoto, "A Proposal of Audience Communications using Video Structures," IPSJ SIG Technical Report 2002-GN-43, pp. 37-42, 2002 (in Japanese).

[16] S. Shimada, K. Miyakawa, and M. Morimoto, "An Analysis on Bulletin Board Communication Synchronized with Video Scene in Fan Community Site," IEICE Technical Report HCS2005-50, pp. 69-74, 2005 (in Japanese).

[17] R. Kataoka, H. Toda, Y. Uematsu, K. Fujimura, K. Bessho, and S. Nishioka, "Navigational Information Retrieval Technologies that Help Users Reach the Desired Information," NTT Technical Review, Vol. 4. No. 8, pp. 17-22, 2006 (this issue).

Masashi Morimoto

Senior Research Engineer, Supervisor, Media Computing Project, NTT Cyber Solutions Laboratories.

He received the B.E. and M.E. degrees in information engineering and the Ph.D. degree in informatics from Kyoto University, Kyoto, in 1986, 1988, and 2006, respectively. He joined NTT Human Interface Laboratories in 1988. He has been engaged in R&D of image, audio, and video processing, multimedia handling, computer vision, and human-computer interaction over networks. From 1996 to 1997, he was a visiting researcher at Stanford University working on image retrieval algorithms. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, the Institute of Image Information and Television Engineers (ITE) of Japan, and the Information Processing Society of Japan.

Shozo Azuma

Research Engineer, Media Computing Project, NTT Cyber Solutions Laboratories.

He received the B.E. degree in information science engineering from Osaka University, Osaka, in 1994 and the M.E. degree in information science engineering from Nara Institute of Science and Technology University, Nara, in 1996. He joined NTT Human Interface Laboratories, Tokyo, in 1996. He has been engaged in R&D of natural language processing, geographical information systems, agent systems, and multimedia processing.

Hiroshi Konishi

Senior Research Engineer, Media Computing Project, NTT Cyber Solutions Laboratories.

He received the B.E. and M.E. degrees in computer science and information mathematics from the University of Electro-Communications, Tokyo, in 1991 and 1993, respectively. He joined NTT Network Information Systems Laboratories, Tokyo, in 1993. From 1995 to 2003, he worked on R&D of e-Learning systems. Since 2003, he has been engaged in R&D of audio and video processing and media-handling technology. He is a member of IEICE and the Acoustical Society of Japan.

Satoshi Shimada

Senior Research Engineer, Media Computing Project, NTT Cyber Solutions Laboratories.

He received the B.E., M.E., and Ph.D. degrees in electrical engineering from Kanazawa University, Ishikawa, in 1984, 1987, and 2004, respectively. He joined NTT Laboratories, Tokyo, in 1987. His research interests include image and video processing, computer vision, and pattern recognition. He is a member of IEICE and ITE.