The t-Room—Toward the Future Phone

Keiji Hirata†, Yasunori Harada, Toshihiro Takada, Shigemi Aoyagi, Yoshinari Shirai, Naomi Yamashita, and Junji Yamato

Abstract

The t-Room is a remote computer supported cooperative work (CSCW) system that we are developing. Our approach is to build rooms with an identical layout, including walls of display screens on which users and physical or virtual objects are all shown at life size, and to provide symmetry of awareness and immersion in each other's physical space. This allows users in the t-Rooms to feel as if they were in the same room. Furthermore, by introducing recording and playback capabilities to the activities that occur within the t-Room, we can achieve asynchronous communication that overcomes the temporal barrier. The t-Room has greater potential to reproduce reality and awareness in remote collaboration than conventional videoconferencing systems. Moreover, as a step toward the future phone, it will enable us to implement a wide variety of telecommunication services that give the feeling of being in the same room and to create a new social interaction style through a large-scale immersive user interface.

1. Introduction

Even with the availability of videoconferencing systems (VCSs), there are still occasions when there is no substitute for direct face-to-face discussions. Conventional VCSs have aimed to create a strong feeling of presence by using high-quality audio and video and virtual reality technology. However, they do not fully recreate the interactions that naturally take place between people who are gathered in the same room. For example, if one person (John) sees another person (Mary) from a diagonal perspective, then Mary should also see John from a diagonal perspective. The apparent distance of John from Mary should also match the apparent distance of Mary from John. If John moves, then his apparent direction and distance from Mary should correspondingly change, and vice versa. Their positional relationship should also be immediately apparent to a third person who is viewing them from the side. These phenomena that people take for granted when they are all in the same room are not conveyed by conventional VCSs. We have developed an advanced telecommunication system called the "t-Room" that creates the feeling of being in the same room for users who are actually separated spatially. To overcome the spatial barrier, we use the simple approach of building rooms with an identical layout, including walls of display screens on which users and physical or virtual objects are all shown at life size.

In addition to overcoming the spatial barrier, we think it is important to support collaboration among people in different time zones, time senses, and time scales. For example, when people in Europe have a meeting at 10 am with people in the USA, the people in the USA have to stay awake until midnight and discuss serious matters while fighting off sleepiness. Arranging a meeting among busy users is always tough, and often people want to postpone participation in a less important meeting because they give higher priority to more pressing tasks. Commonly used asynchronous media, such as email, mailing

† NTT Communication Science Laboratories
Soraku-gun, 619-0237 Japan
Email: hirata@brl.ntt.co.jp

lists, BBSs (bulletin board systems), and blogs (web logs), enable people to contribute to communication at a convenient time, and these media seem to overcome the temporal barrier that exists in VCSs. We have observed that quotations in email and citations in a BBS play a reference role and thus facilitate a kind of interactivity that is different from that in face-to-face communication. The t-Room also overcomes the temporal barrier by enabling a present user to share the awareness information of users in a cited or quoted fragment. We have developed a video editor that can handle the input and output of continuous multistream video and sound data with citation or quotation functions.

To convey the feeling of being in the same room as people who were present at other times, the t-Room allows an earlier conference to be recorded and played back during a later conference. This later conference can itself be recorded and played back during another conference in the future. This playback feature plays a role similar to that of quoted text in email.

This paper is organized as follows. First, we briefly explain the historical background of the t-Room project. Then, we introduce the basic concepts that allow us to feel as if we were actually in the same room. We then describe our ongoing project for building a prototype system that demonstrates this feeling of being in the same place and that explores the possibilities of the concepts. We conclude by mentioning our future plans and our perspectives.

2. Project background

The core members of the t-Room project are Keiji Hirata and Yasunori Harada. Up until April 2003, they had been working independently on various fields of computer science. Hirata was mainly interested in music information processing and built several music systems, such as those for composition, arrangement, and performance rendering [1]. Meanwhile, Harada had also been mainly involved in programming languages and user interfaces and had built many prototype systems. They enjoyed developing systems that they wanted to develop, in the manner of medieval scholars with a patron who could enjoy the free and easy life of scientific curiosity. Fortunately, NTT was very broadminded and gave these two free run of their research fields for more than ten years, for good or for bad. To some extent, they succeeded in achieving satisfactory results outside the company, that is, in domestic and international academic societies. They built experimental systems and wrote interesting technical papers based on the systems and repeated the cycle. The systems were dedicated to their own needs, and the two researchers were the only users of the systems they developed. The lifetime of each system was inevitably very short. Although the two were independently involved in basic research, they came to realize at almost the same time that they would like to do something not only to serve the company and society but also to repay the company's broadmindedness and patronage. They decided that they had to contribute to the telecommunications industry, while meeting the challenges of both business and basic research.

They began to discuss intensively how to launch a new project and arrived at an agreement, which was characterized by the following four points. First, they should develop systems that they could actually continue to use in their daily lives. This may be a reflection of the fact that the lifetimes of the systems developed previously were short. It also implies development of a system where the developer is an enthusiastic user.

Second, they should concentrate on communication science simply because it is the mission of their laboratories. They felt that many of the researchers in the Communication Science Laboratories had been working on research topics related neither to communication science nor the telecommunications industry (although they had to admit that this also applied to them before starting the t-Room project). Hence, they thought that the personnel allocation of the laboratories was out of balance.

Third, they should develop an application that could exchange a huge amount of meaningful, beneficial data over a network. From the viewpoint of NTT's origin, the construction of network devices, which in a sense involves engineering work, seems to be a strong point, whereas providing high-level services and developing novel applications do not seem a specialty. As a result, the constructed network is not used effectively, and there are currently many dark fibers (optical fibers that have been installed to meet future demand but are currently not in use). To make matters worse, a substantial amount of network traffic is worthless, socially harmful, and even illegal data.

Fourth, they should take risks and pursue a new frontier of communication science. They observed that both young and senior researchers throughout the laboratories lacked the spirit to meet challenges, probably because they chose to obtain steady, often trivial, results and just churn out sufficient technical

papers. They both had the experience of changing research fields two or three times and had made many attempts in completely different research fields, even ones that were far from mature. Consequently, they were fortunate in being familiar with working in unexploited, possibly uninhabited research fields.

In summary, Hirata and Harada firmly decided that any research theme considered for the new project should meet the above four criteria. Consequently, they decided to develop a vision-intensive system that prompts social interaction [2]. Admittedly, since the start of the t-Room project, there have been many twists and turns, but we would like to discuss these issues later in another article.

Incidentally, we would like to explain the etymology of "t-Room". Of course, "room" probably needs no explanation, while the meaning of "t" is a communication environment, medium, or something to facilitate communication; it encompasses telephone, time-machine, teleportation, total recall, tea, and tobacco.

3. Basics of the feeling of being in the same room

3.1 Rooms of identical layout

To overcome the spatial barrier, we use the simple approach of building rooms with an identical layout, including walls of display screens on which users and physical or virtual objects are all shown at life size. This enhances the symmetry of awareness (each user acquires the same kind of awareness information) and the mutual immersion in each other's physical space (as many kinds of awareness information as possible are shared among users).

The basic layout of cameras and displays for achieving symmetry and immersion is shown in **Fig. 1**. The technique used in the figure was originally invented for sharing the drawing surface of a display [3]. Since a liquid crystal display (LCD) inherently emits polarized light, the polarizing filter in front of the camera lens effectively cancels out the light from the opposite LCD and captures only the light from real objects in front of the LCD.

We use this technique to share awareness around the surfaces of the LCDs and to make these surfaces into a unified input/output device. Here, the size of the display is crucial to awareness sharing. Since wider displays, in general, can show a larger area of a user's image at life size, they enhance the feeling of immersion in each other's physical space. With the layout in Fig. 1, the two people can, to some extent, recognize gestures and recognize when an object is being pointed to and where it is, as long as the act of pointing is not obscured by other objects. On the other hand, the eye contact recognition provided is still totally inadequate.

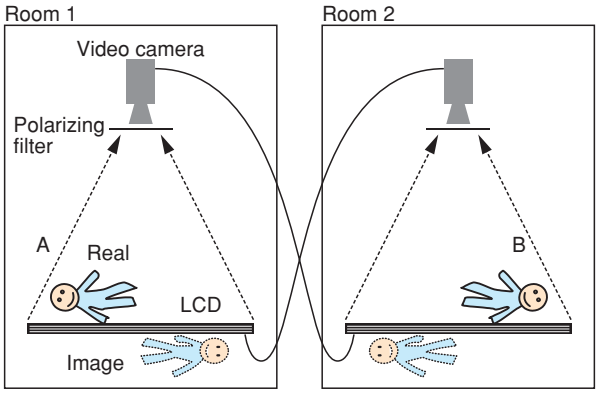

Fig. 1. Basic layout of cameras and displays.

3.2 Playback of recorded activities

To overcome the temporal barrier, we must: 1) enable a present user to share the awareness of users that appear in a cited or quoted fragment and 2) develop an editor that can handle input and output of continuous multistream video and sound data with citation or quotation capability [4]. When we use rooms of an identical layout as unified input/output devices, activities occurring at a remote room, those in the past room, and those in the room where a user is currently present can all be treated in the same way and thus overlapped onto each other.

The other distinctive feature of the t-Room is its metaphor-free structure. When the feeling of being in the same room holds, a user does not need to change his/her behavior after entering the t-Room from that in his/her daily life in the real world. Without using any metaphor, users can understand objects and activities within the t-Room and manage the t-Room workspace in almost the same way as one in the real world.

In this paper, we often used the phrases "feeling of being in the same room" and "feel as if we were actually in the same room" to express the phenomenon provided by a symmetry of awareness and mutual immersion. We think the sense embodied in these phrases can be regarded as the overall criterion for whether a VCS actually succeeds in overcoming spatial and temporal barriers [5].

4. The t-Room system

Based on the above discussion, we aimed to demonstrate and explore the feeling of being in the same room by developing a prototype system, called t-Room [6], as a means to open up a new horizon for communication science. The t-Room system is designed as simply as possible to meet the demands of various styles of group activities that might consist of meetings among several people sitting around a table, a lecture by a professor with many students listening, or a casual meeting such as those that take place in hallways and cafeterias.

4.1 Hardware design

The hardware design of the current t-Room system is shown in **Fig. 2**. A single t-Room consists of six modules (called monoliths) arranged on six sides of an octagon and a worktable at the center containing built-in LCD displays. The 40-inch LCD panels (resolution: 1280 × 768 (WXGA)) should be placed along a polygon to create the feeling of being in the same room. Within the t-Room, LCD panels can be arranged in parallel, while such a layout is impossible in conventional VCSs because the parallel arrangement of LCD panels usually causes visual howling. When the t-Room systems were first built about a year and a half ago, there were seven monolith modules arranged octagonally. However, this number often caused users to feel as if they were locked inside the room, so we removed one. This eliminated the shut-in feeling while completely preserving the feeling of being in the same room.

We installed two identical t-Rooms in the cities of Atsugi and Kyoto, which are approximately 400 km apart. A commercially available 100-Mbit/s optical fiber line (i.e., fiber to the home (FTTH)) connects the two rooms. Currently, video data is transmitted by Motion JPEG (standard set by the Joint Photographic Experts Group) on TCP/IP (transmission control protocol, Internet protocol), and a single video camera transmits data at 200–500 kbytes per second. Audio data is transmitted by PCM (pulse code modulation) on UDP/IP (user datagram protocol, Internet protocol), which requires a bandwidth of about 1 Mbyte/s. The total bandwidth required is about 4–8 Mbyte/s. The actual measured network delay for video data transmission between Atsugi and Kyoto is around 0.7–0.8 s in normal operating conditions. So far, more than 250 people, including visitors, have used the t-Room.

4.2 Gaze

One situation where gaze communication seems to work well is shown in **Fig. 3**. The geometrical relationships observed in Fig. 3 are illustrated in **Fig. 4**, which shows people's positions in gaze communica-

Fig. 3. Gaze communication.

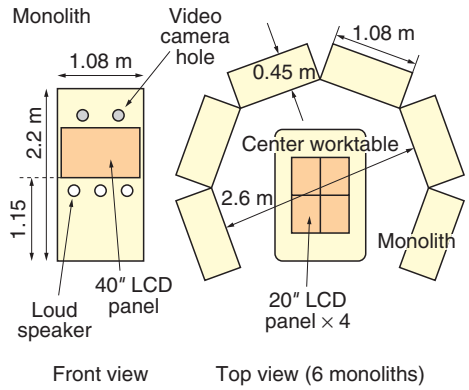

Fig. 2. Hardware design of t-Room system.

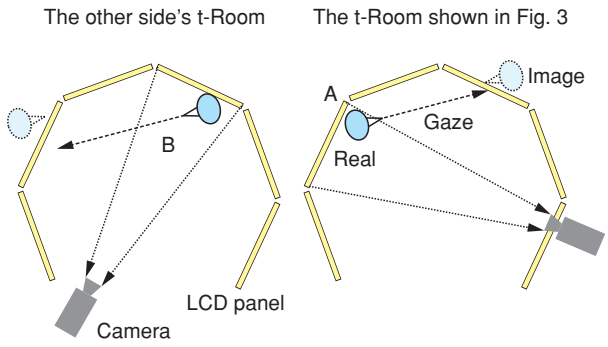

Fig. 4. Geometrical relationships.

tion, gaze directions, and camera angles. In the t-Room on the left-hand side of Fig. 4, person B looks forward and to her right, and the camera captures her image from front-on. Due to the Mona Lisa effect, B's image in the t-Room on the right-hand side looks too far to the side. The Mona Lisa effect is a cognitive phenomenon as follows: when a viewer looks at a person displayed on a screen looking at the front, the viewer feels that the person always appears to follow the viewer with his/her gaze, wherever the viewer stands. The same phenomenon occurs for A's image as seen by B.

Moreover, since the video cameras are positioned just above the 40-inch LCD panels, they cannot avoid capturing images of users in front of the opposite LCD panels at a downward angle (2.6 m ahead and about 25 cm down, i.e., a gradient of about –0.1), which also degrades gaze communication [7]. However, even if a camera could be placed just between two LCD panels that were next to each other at the middle height, occlusion would occur due to users walking around inside the t-Room. We think that this type of occlusion is more problematic, so we placed the cameras above the LCDs as close to their upper side as possible.

4.3 Recording and playback

4.3.1 Sharing awareness

A snapshot of a playback in which three t-Room spaces are overlapped is shown in **Fig. 5**. Persons A and C are present, while person B was present eight months ago: A is present at location 1, while B and C were/are at location 2, respectively. Here, A and C first look at B's playback image and then look in the same direction as B to discover what B saw at the time of recording. Since the t-Room always records all of the activities that occur within it, A and C can discover the object that B was looking at by following the end point of B's gaze. As a result, A and C successfully share B's workspace awareness.

The action of C pointing somewhere at B's image at a certain time during playback plays the role of a citation. That is, the capability of spatial and temporal citing of video and audio data is achieved in a quite natural way by using the t-Room as a unified input/output device. For example, at a certain moment, person C pauses the playback of B's action and adds some comments to what B just said. Of course, person A can do this as well. Note that in the t-Room, actions by A and C, such as pausing and adding, are also all recorded, which facilitates further citation and quotation by someone else. Unfortunately, the current implementation can only support playback with one level of citation.

The current t-Room system fixes the order of overlapping images and the alpha channel value (transparency ratio); present images with an alpha channel value of 0.5 come in front of those in the past. Therefore, users in the present see partially transparent recorded images. The order of overlapping and the individual alpha channel values may be changed, depending on the style of activity and/or demand.

4.3.2 Snapshot

If the local person B points at a component of the personal computer (PC) on the worktable, remote person A can naturally see what B is pointing at (Case 1 in **Fig. 6**). However, in the reverse situation (Case 2

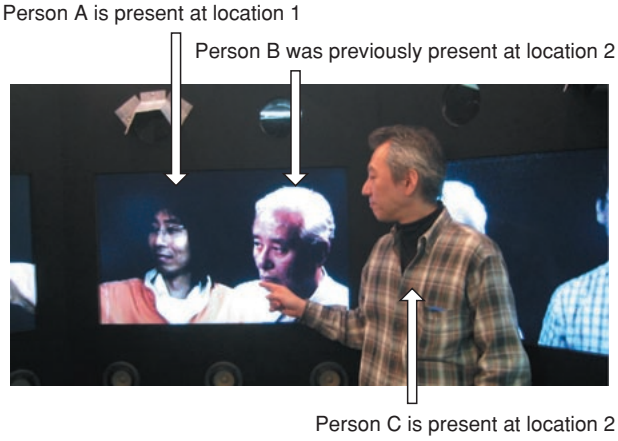

Fig. 5. Playback with recorded and remote scenes.

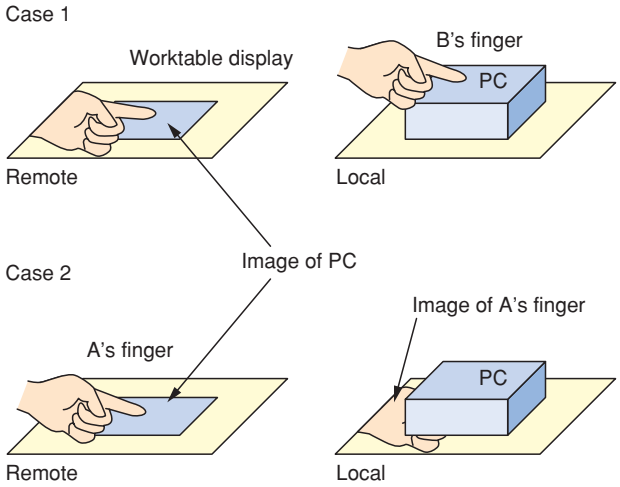

Fig. 6. PC hides the remote user's finger.

in Fig. 6), in which A points at a component in an image displayed on the remote worktable, B cannot see what A is pointing at because the scene on the remote worktable is correctly displayed at life size on the local worktable, so the PC on the local worktable hides the displayed image.

Through practical usage over several months, we found that the snapshot capability was useful for coping with this difficulty. Snapshots are also regarded as a special case of a playback (i.e., instantaneous playback). A scene with a snapshot shared by two people, A and B, is shown in **Fig. 7**. A snapshot of a broken PC placed on the center worktable is taken by a local camera and displayed on identical LCD panels in both the local and remote t-Rooms. The two people in the different t-Rooms can correctly recognize which component the other person is referring to due to the symmetry of awareness along with the feeling of being in the same room.

5. Preliminary experiment

5.1 Outline

As a step towards demonstrating the feeling of being in the same room, we conducted a preliminary experiment to examine the advantages and disadvantages of the t-Room compared with a conventional VCS. Eleven adults in their 20s and 30s participated; all of them were novice PC users who had performed simple tasks with a PC but had never repaired one. In addition, a subject who played the role of an instructor participated; he had been a lecturer in PC usage at a technical college for two years. In turn, each novice PC user in the local t-Room was asked to repair a broken PC with the help of the instructor in a remote t-Room. The instructor did not watch. The pair of instructor and repairer was assigned three subtasks: replace the power supply unit, the hard disk drive, and optical disk drive. Three conditions were set up as described below.

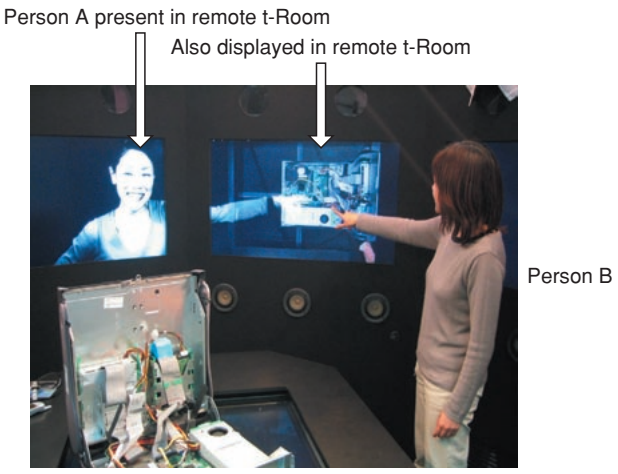

Fig. 7. Pointing at component on snapshot.

C1 (Conventional VCS): The repairer sees the instructor shown on the opposite LCD panel in front of him/her, while the instructor can see only the image captured by a WebCam that the repairer holds in his/her hand.

C2 (t-Room with self-reflection image): This condition corresponds to a basic function level of the t-Room, which is considered a baseline.

C3 (C2 and snapshot): We introduced the snapshot because it is a special case of recording and playback capability. Since an experiment properly designed for recording and playback capability would take a lot of time, we focused on a short time period and used a snapshot.

5.2 Results (feedback from participants)

All repairers successfully complete the three subtasks, taking from 20 to 70 minutes per task. Overall, 9 out of the 11 repairers reported that C3 was the easiest of the three conditions to work in, and the other 2 said C1 was. 10 of the repairers said that the easiest subtask was replacing the power supply unit followed by the hard disk drive and optical disk drive.

In C1, repairers frequently mentioned the usability of a handy WebCam and whether or not repairers could understand the instructor (T)'s instructions. One repairer (R1) found C2 easier because R1 and a PC on the worktable were entirely captured by surrounding cameras from a fixed viewpoint, and there was no need for R1 to control the viewpoint. Another repairer (R2) pointed out that in C2, it was difficult to see T's finger shown on the worktable and hence to determine what it pointed at. R2 also said that the gap between T's fingertip and the target being pointed at by it just confused him. In C3, five repairers reported that using a snapshot improved certainty; it was like using a blackboard. The snapshot enabled R3 to easily understand what T was saying. R4 was happy that the snapshot enabled T to understand what he said and vice versa.

6. Conclusion

The t-Room is a remote collaboration system that provides symmetry of awareness and immersion in each other's physical space that allows users in sepa-

rate t-Rooms to feel as if they were actually in the same room, unlike a conventional videoconferencing system. In launching the t-Room project, we had four goals: to improve our daily lives, contribute to communication science, exchange a huge amount of meaningful, beneficial data, and take risks in basic research. We think that these four points are still relevant and that the t-Room project could lead to their successful achievement.

Provisional results for a preliminary experiment with the t-Room, in which a novice user repaired a PC while consulting a remote instructor, imply that the t-Room has more potential to reproduce reality and awareness in remote collaborations (the feeling of being in the same room) than is possible with conventional videoconferencing systems.

We think that future work should include developing full-fledged recording and playback facilities, providing sound capability that supports the feeling of being in the same room, building a large, multi-location t-Room, conducting further experiments and evaluations, and designing and implementing flexible, scalable middleware for a variety of applications.

We demonstrated this system at NTT Communication Science Laboratories Open House 2005 (June 2005), and many visitors were highly impressed by the feeling of being in the same room. In December 2005, we launched the official "t-Room" website [8]. This site is packed with all sorts of information, including a description of the technology behind t-Room and an 11-minute demonstration video. You are invited to take a look.

Finally, we do not want readers to underestimate the potential of the t-Room. We are developing the t-Room as a future form of telephone service. Toward the future phone, it will be possible to provide a wide variety of communication services while giving the feeling of being in the same room with people at different places and different times, thus creating a new social interaction style with large displays. In this sense, we think that calling our system a "videoconferencing system" would be inappropriate because that would give people a false mental image that fails to convey its futuristic abilities. We prefer to avoid using that term.

References

[1] K. Hirata, "Musical Knowledge Programming for a New Form of Music Distribution," NTT Technical Review, Vol. 2, No. 12, pp. 6-11, 2004.

[2] D. A. Norman, "Emotional Design: Why We Love (or Hate) Everyday Things," Basic Books, 2005.

[3] J. C. Tang and S. L. Minneman, "VideoDraw: A Video Interface for Collaborative Drawing," Proc. of CHI '90, pp. 313-320.

[4] T. Takada and Y. Harada, "Citation-Capable Video Messages: Overcoming the Time Differences without Losing Interactivity," Proc. of Information Spaces and Visual Interfaces 2000, pp. 31-38.

[5] Y. Harada, "Communication with the feeling of being in the same room," Proc. of Workshop on Interactive System and Software '98 (in Japanese).

[6] K. Hirata, Y. Harada, T. Ohno, T. Yamada, J. Yamato, and Y. Yanagisawa, "t-Room: Telecollaborative Room for Everyday Interaction," Proc. of The 66th IPSJ Annual Convention, 4B-3, 2004.

[7] L. Mühlenbach, M. Böcker, and A. Prussog, "Telepresence in Videocommunications: A Study on Stereoscopy and Individual Eye Contact," Human Factors, Vol. 37(2), pp. 290-305, 1995.

[8] www.mirainodenwa.com (in Japanese).

Keiji Hirata
Senior Research Engineer, Computer Scientist, Principal Engineer of Media Interaction Principle Open Laboratory, NTT Communication Science Laboratories.

He received the B.E. degree in metal engineering, M.E. degree in information engineering, and D.Eng. degree in information engineering from the University of Tokyo, Tokyo, in 1981, 1983, and 1987, respectively. He joined NTT Basic Research Laboratories in 1987. He spent 1990 to 1993 at the Institute for New Generation Computer Technology (ICOT), where he engaged in R&D of parallel inference machines. In 1999, he moved to NTT Communication Science Laboratories. He was promoted to Distinguished Researcher in 2001. His research interests include musical knowledge programming and human-computer interaction. He received the Takahashi Award for a paper presented at the annual convention of the Japan Society for Software Science and Technology in 1987 and the IPSJ Best Paper Award from the Information Processing Society of Japan (IPSJ) in 2001. He co-edited and wrote the book chapter "Musical Knowledge Representation on a Computer" in the book "The World of Computers and Music: Foundations to Frontiers" (in Japanese) in 1998 and co-translated the book "Computer Music Tutorial" by Curtis Roads in 2001. He is a member of IPSJ, the Japanese Society for Artificial Intelligence, and the Japan Society for Software Science and Technology (JSSST).

Yasunori Harada
Senior Research Engineer, Chief Technology Officer of Media Interaction Principle Open Laboratory, NTT Communication Science Laboratories.

He received the B.S. degree in applied physics, M.E. degree in information engineering, and Dr.Eng degree in information engineering from Hokkaido University, Hokkaido, in 1987, 1989, and 1992, respectively. He joined NTT Basic Research Laboratories in 1992. He spent 1998 to 2001 at PRESTO-JST (Precursory Research for Embryonic Science and Technology, Japan Science and Technology Agency). His research interests include visual language and object-oriented programming.

Toshihiro Takada
Senior Researcher, Media Interaction Principle Open Laboratory, NTT Communication Science Laboratories.

He received the B.E. and M.E. degrees in information science from Tokyo Institute of Technology, Tokyo, in 1986 and 1988, respectively. He joined NTT in 1988. His research interests are in networked information systems, networked computation, real-space computing, and human-computer interaction. He is a member of the Association for Computing Machinery (ACM), JPSJ, JSSST, and the Human Interface Society of Japan.

Shigemi Aoyagi
Research Scientist, Media Interaction Principle Open Laboratory, NTT Communication Science Laboratories.

He received the B.E. and M.E. degrees in information science from Tokyo Institute of Technology, Tokyo, in 1988 and 1990, respectively. He joined NTT in 1990. His current research interests include parallel processing, distributed algorithms, image understanding, object recognition, and distributed systems for multimedia content. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, IPSJ, and JSSST.

Yoshinari Shirai
Researcher, Ambient Semantics Research Group, Innovative Communication Laboratory, NTT Communication Science Laboratories.

He received the B.A. degree in environmental information and M.E. degree in media and governance from Keio University, Tokyo, in 1998 and 2000, respectively. He joined NTT Communication Science Laboratories in 2000. His research interests include ubiquitous computing and interaction design. He is a member of ACM, IPSJ and the Human Interface Society of Japan.

Naomi Yamashita
Ambient Semantics Research Group, Innovative Communication Laboratory, NTT Communication Science Laboratories.

She received the B.E. and M.I. degrees in applied mathematics and physics and the Ph.D. degree in informatics from Kyoto University, Kyoto, in 1999, 2001, and 2006, respectively. She joined NTT in 2001. She is interested in computer supported collaborative work and interaction analysis.

Junji Yamato
Senior Research Scientist, Supervisor, Group Leader of Recognition Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received the B.E, M.E., and Ph.D. degrees in precision machinery engineering from the University of Tokyo, Tokyo, in 1988, 1990, and 2001, respectively, and the S.M. degree in electrical engineering and computer science from the Massachusetts Institute of Technology in 1998. He joined NTT Human Interface Laboratories in 1990. His research interests are in machine computer vision, human-computer interaction, and robotics. He is a member of IEEE, ACM, IEICE, and the Robotics Society of Japan.