S-TEL: AN AVATAR BASED SIGN LANGUAGE TELECOMMUNICATION SYSTEM

Tomohiro Kuroda, Kosuke Sato,
and Kunihiro Chihara
Nara Institute of Science and Technology, Japan

ABSTRACT

Although modern telecommunication has drastically our daily lives, the audibly challenged cannot benefit from such telecommunication because it is based on phonetic media. This paper introduces a new telecommunication system for sign language utilizing VR technology, which enables natural sign conversation on the conventional analogue telephone line. In this method, a person converses with his/her party's avatar instead of the party's live video. As a speaker's actions are transmitted as kinematic data, the transmitted data is ideally compressed without losing both the language and non-language information of spoken signs. A prototype system, S-TEL, which implements this method on UDP/IP, proved the effectiveness of avatar-based communication for sign conversation via a real lossy channel.

1. Introduction

Although modern telecommunication has drastically changed our social communication style, the audibly challenged cannot benefit from such telecommunication because it is based on phonetic media. A new telecommunication system for signers is indispensable to make their community less isolated and to enrich the quality of their daily lives.

Today the audibly challenged use TTY or facsimiles instead of telephones. However, with these character based communication systems, they need to translate their conversation into a descriptive language and to write down or type in such language. Therefore, a new telecommunication system that enables them to talk in signs is indispensable.

Currently, much research focusing on computer aid for signers including telecommunication systems for signers are coming. Some of such research has developed data compression techniques for the video stream of sign language (Ohki, 1995), while other research has developed script-based sign communication methods that translate signs into descriptive languages to reduce transmitted data (Gulska, 1990). These methods succeeded in compressing transmitted data, but the language and non-language information contained in signs is lost due to the above compression or the translation. Thus, the foregoing systems based on these methods cannot mediate natural sign conversation. Therefore, Kuroda et al. (1995) has introduced a concept of a new telecommunication method for sign language integrating human motion sensing and virtual reality techniques. In this paper, a new realized telecommunication system based on this method is introduced.

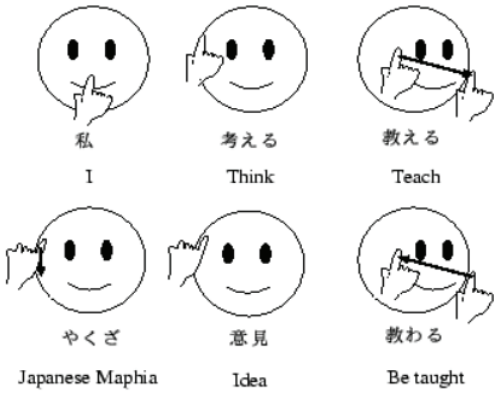
In this system, a person converses with his/her party's avatar instead of the party's live video. Speaker's actions are obtained as geometric data in 3D space, the obtained motion parameters of the actions are transmitted to the receiver, and the speaker's virtual avatar appears on the receiver's display. Thus, it realizes optimal data compression without losing language or non-language information of given signs. Moreover, the speakers can hide their private information without displeasing the receivers.

In this paper, the features of Japanese Sign Language are briefly explained in section 2 and foregoing studies on sign communication are mentioned in section 3. In section 4, the avatar-based communication method is introduced, and avatar-based communication and video-based communication is compared from a bandwidth viewpoint. Finally, in section 5, a prototype avatar-based communication system, S-TEL, is discussed in experiments on UDP/IP by Deaf people and sign experts.

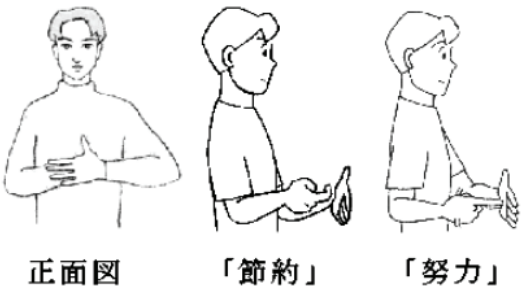
2. Japanese Sign Language

Japanese Sign Language (JSL) has been the mother tongue of Deaf people in Japan for a hundred years. As JSL is a visual language, there are some features in comparison with phonetic languages:

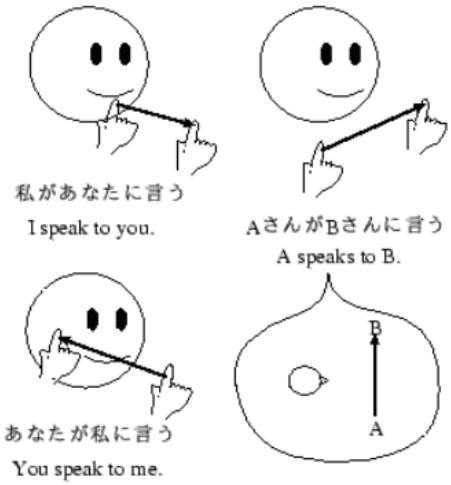
- The meanings of signs are defined by the feature, position, movement, and direction of the hands (Fig. 1-A).
- Some signs cannot be identified from a frontal view because of occlusions (Fig. 1-B).
- Three-dimensional position of signs denotes the relation of the current communication. Persons and their social relations are shown by the position where the signs are presented or the directions of upper body (Fig. 1-C).
- Non-Manual Signals (NMS) like nodding work as modifiers.



(a) The meanings are determine By hands figure etc.



(b) Some signs cannot be identified from a frontal view because of occlusions.



(c) The position of signs shows character and his/her rank.

Figure 1: Features of Japanese Sign Language

3. Foregoing Research

There are many research efforts on computer aid for signers. However, most of them focus on sign translation or sign education. Only a few attempts have so far been made at communication among signers. The research on communication among signers can be divided into two groups: script-based communication and video-based communication.

3.1 SCRIPT-BASED COMMUNICATION

Jun et al. (1991) and Ohki (1995) have proposed script-based communication systems. These systems translate given signs into script language or phonetic language, transmit them, and produce sign animation on the receivers' terminal. Therefore, those systems can drastically eliminate bandwidth of transmission. However, as these script-based systems have a translation stage, they cannot forward given signs when their dictionaries have no entry for the given signs or the systems mistranslate it.

3.2 VIDEO-BASED COMMUNICATION

As sign language is a visual language, a videophone seems usable as a telecommunication system for sign language. However, Yasuda (1989) pointed out the following 'Eyes Torture' problem of videophones:

- User's eyes may break into the other's private space through camera. The other's privacy may be trespassed.
- Kamata (1993) experimented with videophones. The experimental results show the following problems which make sign conversation difficult on videophones:
- 2D videophone images cannot show whole sign information, because sign language moves in 3D space.
- The received images do not have enough resolution, view, and frame rate for practical sign conversation.
- As sign language includes fast hand motions and occluded postures, general methods for video compression are not suitable for sign communication. Sperling et al. (1985) and Gulska (1990) proposed video compression methods for sign language, but their methods cannot transmit sign image sequences that have enough time/space resolution to read on the conventional analogue telephone line.

3.3 BANDWIDTH OF VIDEO-BASED COMMUNICATION

Kamata (1993) argues that the image sequence of QCIF mode ISDN videophones (176 x 144 pixel x 15 fps) doesn't always have enough time/space resolution for sign conversation. However, Sperling et al. (1985) argues that a three bits gray scale image sequence of 24 x 16 (pixels) x 15 (fps) can visualize 'enough intelligible' ASL. Assuming a three bits gray scale image, Kamata's system requires at least 2.2Mbps and Sperling's requires 34Kbps for bi-directional communication. As these results vary widely, we examined required bandwidth for video-based sign communication.

First, to examine required frame rate, we selected two topics (79 seconds) from NHK sign language news, and measured how long the frames for each sign word continued.

As Table 1 and Fig. 2 show, some words continue less than 1/30 second. Moreover, Kanda et al. (1996) discussed that newscasters speak about 70% speed of normal sign conversation. Thus, video rate (30 frames per second (fps)) is not sufficient for sign conversation. However as there are no faster display for home use, we assume video rate display in the following discussion.

Table 1. *The Number of Frames of Each Sign. (1^- denotes less than 1 frame.)*

	News 1	News 2	Total
Average	5.72	9.36	7.84
Minimum	1^-	1^-	1^-
Maximum	21	30	30

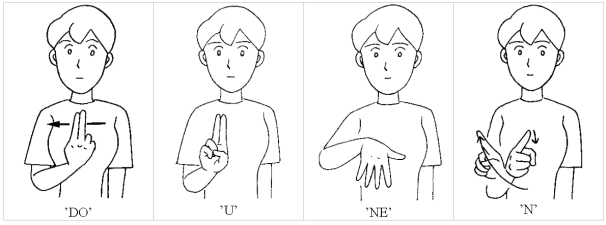


Figure 2: Finger Spelling 'DOUNEN' (Power Reactor and Nuclear Fuel Development Corporation) from NHK Sign Language News. Finger Character 'NE' continues less than 1/30 second.

Second, to examine required resolution, we selected 10 frames (640 x 480 pixels) from NHK sign language news, and measured the width of fingers. The most narrow finger (the pinkie of woman) width was five pixels. Thus, the sign image requires 128 x 96 pixels to identify each finger.

From these discussions, assuming a three bits gray-scale image, the required bandwidth of bi-directional video-based sign communication is 2.2Mbps. Therefore, video-based sign communication requires a broadband digital channel.

4. System Design

4.1 AVATAR-BASED COMMUNICATION

To solve the above problems of the previous telecommunication systems, we introduce a new telecommunication system for sign language integrating both human motion sensing and virtual reality techniques. This method solves natural sign conversation on a conventional analogue telephone line.

In this method, a person converses with his/her party's avatar instead of the party's live video. Speakers actions are obtained as geometric data in 3D space; the obtained motion parameters of actions are transmitted to the receiver, and the speaker's virtual avatar appears on the receiver's display. This avatar-based communication has the following advantages: First, sending kinematic data without any translation process, this avatar-based communication realizes data compression without losing the language or non-language information of given signs. Second, sending 3D motions, the receiver's terminal can produce signing avatar to increase the readability of signs. Third, this method can be applied to conferencing or party talking (Kuroda, 1997b). Finally, visualizing the avatar instead of live video, users can hide their private information without displeasing his/her party.

As shown in Fig. 3, this system consists of following components:

- The **sender** obtains signs as geometric data and sends it. The **motion measuring part** measures hands, arms and upper body motions. The **encoder** encodes and compresses obtained data.
- The **receiver** receives kinematic data of signs and displays avatar. The **decoder** decodes given data into kinematic data. The **avatar producing part** produces avatar from given kinematic data and displays it. This part makes use of the reader's viewpoint information if needed. The avatar can be virtual CG avatar, robot, etc.

4.2 BANDWIDTH OF AVATAR-BASED COMMUNICATION

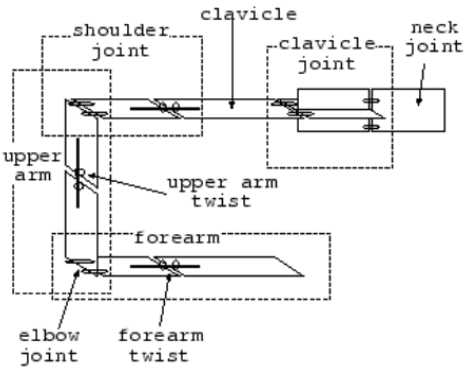
The avatar must handle the whole upper body as a signer shows signs with hand, arm and upper body motions. Therefore, we assume a skeleton of the avatar as shown in Fig. 4. This model has the following features.

- The spine consists of 33 or 34 vertebras, and small rotation between these vertebras produces a backbone bend. Especially, five cervical vertebras and seven lumber vertebras moves much more than the other vertebras. Thus, this model has joints at both sides of

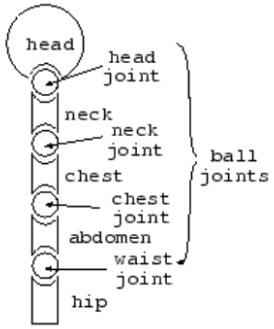
Figure 3: Overview of Avatar-based Communication

these two parts.

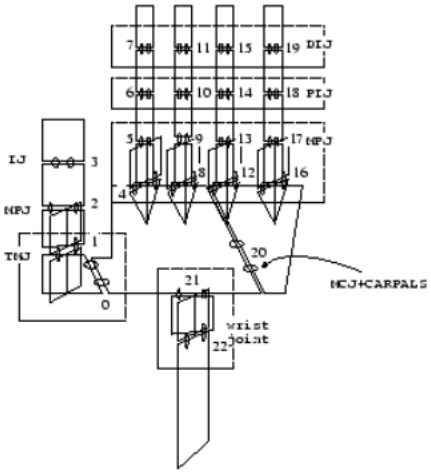
- Some sing words like ASL 'why' visualized by shoulder motion. Therefore, this model has joints on the neck side of the clavicle.



(a) Arm model



(b) Body model



(c) Hand model

Figure 4: Skeleton Model of Avatar

Suppose finger width is unit length as in section 4.3. Finger width for a woman is about 2 cm and the sitting woman's wrist moves inside a sphere whose diameter is 2m. Therefore, seven bits code can denote wrist position, and eight bits code can denote 360^o . This model has 72 degrees of freedom and the amount of rotation of each joint is as in Table 2. Thus, 460 bits code can denote whole upper-body posture. Kuroda et al. (1996b) proposed a method to reconstruct upper body motion form position and rotation data of wrists and the top of the head. Applying this method, 415 bits code can denote whole upper-body posture. From this discussion, the required bandwidth of bi-directional avatar-based sign communication is 16Kbps. Thus, avatar-based communication is available on a conventional analogue telephone line.

Table 2. *Joints of Avatar*

Part	Degrees of freedom	Rotation	Bits
Hand	42	90	6
Wrist	4	180	7
Clavicle joint	4	90	6
Rotation around clavicle	2	270	8
Others	20	180	7

5. Experiments

5.1 PROTOTYPE SYSTEM S-TEL

A prototype, S-TEL, along the design discussed in section 4, is developed as in Fig. 5. Kuroda et al. (1996a) showed that a 3D stereo scopic view has no effect on the readability of signs, and that a 2D CG reflecting readers motion parallax is sufficient to realize practical readability. Therefore, S-TEL uses normal 2D display as shown in Fig. 6.

S-TEL sender is composed of a Pentium 166MHz PC with Windows95, two CyberGloves and a Fastrak. S-TEL receiver is composed of Intergraph TD-5Z workstation (Pentium 100MHz with OpenGL accelerator) with WindowsNT 3.51 and a Fastrak. All software components are built on World Tool Kit Ver. 2.1 for WindowsNT and Visual C++ 2.

5.2 BANDWIDTH OF S-TEL

S-TEL obtains the signer's action with CyberGloves and Fastrak. CyberGlove obtains 18 finger-joint bending as integers and Fastrak obtains position as three single floats and orientation as a single float quaternion (4 degrees of freedom). Therefore, the amount of data of one frame is 120 bytes. Assuming 30 fps, the bandwidth of mono-directional S-TEL is 28.8Kbps.

5.3 EXPERIMENTS

We experimented with S-TEL on two types of UDP/IP communication channels. First, we connected S-TEL by the

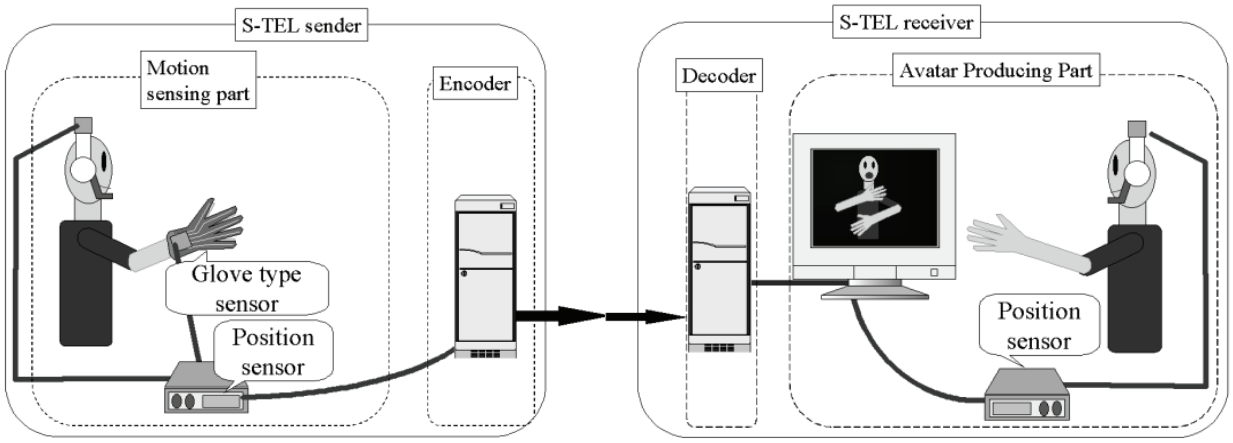


Figure 5: Design Overview of S-TEL

Ethernet. Three Deaf people, three sign experts, and four sign beginners tried to talk freely in sign language on S-TEL.

Second, we placed S-TEL sender at NAIST, Ikoma City (Nara) and S-TEL receiver at Kumamoto City (Kuroda et al., 1997a). The distance between the two cities is about 700Km. Satellite JCSAT-1 connected two cities. The bandwidth of the channel was 4.0Mbps bi-directional. By adding heavy traffic on the communication channel, there is 2% packet loss due to overload of the channel. The speakers were two sign experts and readers were one sign expert and two beginners. All testees are hearing. Testees tried to teach sign language on S-TEL. Testees also tried to talk in signs through NV/VAT Internet video chat, and they compared these two systems.

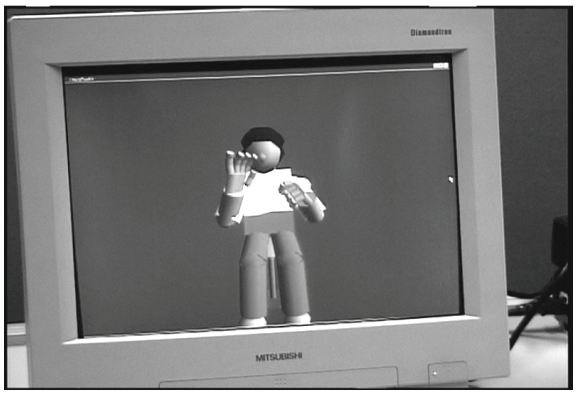


Figure 6: Produced Avatar of S-TEL

5.4 RESULTS AND DISCUSSIONS

Experimental results are as follows.

- The realized frame rate of S-TEL is 26.1 fps. S-TEL can perform a practical frame rate to read signs on a consumer PC.
- 2% packet loss had no effect on the quality of visualized signs, because S-TEL detects erroneous frames and keeps the continuity of the preceding frame images. This ensures S-TEL works properly on lossy channel.
- Readers could recognize 75% of spoken signs. Users can almost make themselves understood through S-TEL. Speakers signs which readers couldn't recognize were the following cases:
- Readers couldn't recognize spoken signs which facial expressions modify, because S-TEL doesn't treat facial expressions.
- Readers couldn't recognize finger spelling, because the finger size of the avatar is not sufficient to distinguish fingers.
- Sign experts completed a difficult task to teach 'self-introduction' to beginners over S-TEL. Usually, for a beginner, one must teach to a beginner's face with utmost care and kindness. It is one proof that S-TEL can provide the environment where users can talk in signs as if they talk face to face.
- Testees said that they prefer S-TEL rather than videophones as a sign communication media because videophones exposed their privacy, but S-TEL didn't. They also said S-TEL would enrich their daily lives.

These experimental results show that avatar-based sign communication is effective because users can make themselves understood through S-TEL. Moreover, avatar-based communication is superior to video-based system in bandwidth viewpoint; avatar-base communication is available on lossy and narrow channel.

To increase the readability of signs, avatar-based communication should treat facial expressions and visualize the hands larger. However, to make the hands larger causes other problems. We tried to make the hands larger, but testees complained that unbalanced avatar makes signs unreadable and that testees sometimes feel hit by the avatar

when the avatar stretches their arm to the front. Thus, to increase readability, avatars hands should expand when the speaker starts to spell. A new method to distinguish finger spelling should be developed.

On the other hand, visualizing facial expression on avatar based communication is easily realized by attaching a texture of the face image on the avatar as shown in Fig. 7. However, bi-directional face image transmission needs an ISDN channel. Moreover, texture replacement requires much CPU power. Thus, the system should utilize the model based facial expression encoding (Harashima et al., 1989) and visualize the face using CG primitives.

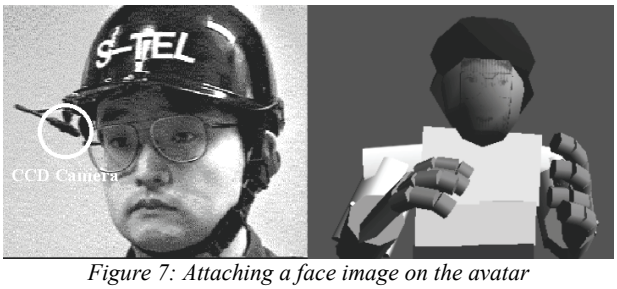


Figure 7: Attaching a face image on the avatar

6. Summary

In this paper, an avatar-based sign communication system, which is an innovative system for sign language telecommunication, is presented. In this method, speakers actions are obtained as geometric data in 3D space, the obtained motion parameters of actions are transmitted to the receiver, and the speaker's virtual avatar appears on the receiver's display. As avatar-based communication treats the speaker's actions as geometric data in 3D space, it allows the users to talk and read signs naturally through a conventional analogue telephone line, increase readability of signs, and protects users' privacy.

Experiments to talk in signs through a prototype system, S-TEL, were performed. The results showed the effectiveness of avatar-based communication and the superiority of avatar-based communication over video-based communication as a telecommunication media for sign language.

When S-TEL becomes popular among Deaf people, their community should become less isolated. S-TEL should increase the quality of their daily lives. We are currently integrating the intelligent transmission of speaker's facial expressions and the identification of finger spelling into S-TEL.

7. Acknowledgement

This study is in cooperation with Kamigyo Branch of Kyoto City Sign Language Club "Mimizuku", and Kamigyo Branch of Kyoto City Federation of Deaf, Wide Project, Digital Research Inc. and Japan Satellite Systems Corp.

REFERENCES

S Gulska (1990), The Development of a Visual Telephone for the Deaf: Using Transputers for Real-time Image Processing, *In Transputer Research and Applications 3. Proceedings of the Third North American Transputer Users Group*, pp.7-16

H Harashima, K Aizawa and T Saito (1989), Model-based Analysis Synthesis Coding of Videoteleophone Images: Conception and Basic Study of intelligent Image Coding, IEICE transactions, **72**, 5, pp.452-459

X Jun, Y Aoki and Z Zheng (1991), Development of CG System for Intelligent Communication of Sign Language Images between Japan and China, IEICE transactions, **74**, 12, pp.3959-3961

K Kamata (1993), Sign Language Conversation through Video-phone – Experiment at School for the Deaf –, Technical Report of IEICE, ET92-104, pp.37-44, Japanese

K Kanda (1996), Sign Linguistic from Basics, Fukumura Press, Japanese

A Komekawa (1998), Japanese-Sign Language Dictionary, Japan Federation of Deaf, Japanese

T Kuroda, K Sato and K Chihara (1995), System Configuration of 3D Visual Telecommunication in Sign Language, *In Proceedings of the 39th Annual Conference of the Institute of Systems, Control and Information Engineers, ISCIE*, pp.309-310, Japanese

T Kuroda, K Sato and K Chihara (1996a), S-TEL: A Telecommunication System for Sign Language, *In Conference Companion of First Asia Pacific Computer Human Interaction*, pp.83--91

T Kuroda, K Sato and K Chihara (1996b), Reconstruction of Signer's Actions in a VR Telecommunication System for Sign Language, *In Proceedings of International Conference on Virtual Systems and Multimedia VSMM'96 in Gifu*, pp.429-432

T Kuroda, K Sato and K Chihara (1997a), S-TEL: A Sign Language Telephone using Virtual Reality Technologies, *In CSUN's 12th Annual Conference Technology and Persons with Disabilities*, Floppy Proceedings, KURODA_T.TXT

T Kuroda, K Sato and K Chihara (1997b), S-TEL: VR-based Sign Language Telecommunication System, *In Abridged Proceedings of 7th International Conference on Human-Computer Interaction*, pp.1-4

M Ohki (1995), The Sign Language Telephone, TELECOM '95, pp.391--395

G Sperling, M Landy, Y Cohen and M Pavel (1985), Intelligible Encoding of ASL Image Sequence at Extremely Low Information Rates, Computer Vision, Graphics, and Image Processing, **31**, 3, pp.335-391

H Yasuda (1988), TV-phone Now, Spectrum, **1**, 5, pp.88-102, Japanese

BIOGRAPHIES

Tomohiro Kuroda received B.S. in information science from Kyoto University, Japan in 1994, M.S. and Ph.D. in information science from Nara Institute of Science and Technology, Japan in 1996 and 1998. Since 1998, he has been an instructor at Nara Institute of Science and Technology. He has been a technical trainee at the Biomedical Technology Laboratory, Helsinki University of Technology, Finland, in 1995. His current research interests include Sign Linguistic Engineering, Assistive Technology and Virtual and Augmented Reality.

Contact information:
Tomohiro Kuroda
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0101, Japan
Email: tomo@is.aist-nara.ac.jp

Kosuke Sato received his B.S., M.S. and Ph.D. degrees in control engineering from Osaka University, Japan in 1983, 1985 and 1988, respectively. He was with the Department of Control Engineering at Osaka University from 1986 to 1994. Since 1994, he has been with Graduate School of Information Science, Nara Institute of Science and Technology, where he holds the rank of Associate Professor. He has been a visiting scientist at the Robotics Institute, Carnegie-Mellon University, PA, from 1988 to 1990. His current interests are in 3D range sensing for Computer Vision, Augmented Reality, and Human-Computer Interaction.

Contact information:
Kosuke Sato
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0101, Japan
Email: sato@is.aist-nara.ac.jp

Kunihiro Chihara received his B.S., M.S. and Ph.D. degrees in control engineering from Osaka University, Japan in 1968, 1970 and 1973, respectively. He was with the Department of Control Engineering at Osaka University from 1973 to 1991. Since 1992, he has been a Professor at Graduate School of Information Science, Nara Institute of Science and Technology. He was the director of Information Technology Center from 1994 to 1998. He is currently the director of Research Center for Advanced Science and Technology from 1998. His current interests are in Virtual Reality for Medical applications and Multimedia Information Processing.

Contact information:
Kunihiro Chihara
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0101, Japan
Email: chihara@is.aist-nara.ac.jp