Designing a video-mediated collaboration system based on a body metaphor

Hiroshi Kato1), Keiichi Yamazaki2), Hideyuki Suzuki1), Hideaki Kuzuoka3), Hiroyuki Miki4), and Akiko Yamazaki5)

1) C&C Media Research Laboratories, NEC Corporation, 2)Faculty of Liberal Arts, Saitama University, 3)Institute of Engineering Mechanics, University of Tsukuba, 4)Media Laboratories, Oki Electric Industry Co. Ltd., and 5)Nursing School of Douai Hospital

Abstract

Optimum arrangement of communication resources for a distance education system to support collaborative learning, including hand-manipulation of physical objects, was empirically investigated. The authors propose a 'body metaphor' concept, which follows the ordinary body arrangement in everyday instruction. It allows: (1) to display instructor's pointer; (2) to display instructor's face; (3) to display views of learner's orientation; (4) to arouse learner's awareness of being watched by the instructor; and (5) to dispose communication resources apart enough so as to clarify the learner's orientation.

Comparative experiments revealed that body metaphor setting supported smoother collaboration than the conventional face-to-face metaphor setting.

Keywords— distance education, CSCW, collaborative learning, ethnomethodology, interaction analysis, video conference

Introduction

Most of the distance education systems in use today have been built on the basis of a bi-directional video-mediated tele-conference system, i.e., views of a lecturer's face and/or learning material are transmitted to the learners' sites, and vice versa. In such a system, however, it has been pointed out that gaze, gestures, and other body movements are generally not as effective as in normal face-to-face communication _iHeath et al., 1991, 1992). Therefore, the prevailing systems are not sufficient, particularly for conducting collaborative learning in which non-verbal communication plays an important role, such as in a scientific experiment or in a physical exercise, although they may be acceptable for an education style like lectures in which symbolic (verbal or visual) information transmission from a teacher to learners is dominant.

Research in computer supported cooperative work (CSCW) has studied this problem mainly in regard to transmission of hand gestures and the question of eye contact. Some CSCW systems, such as VideoDraw (Tang et al., 1990) and ClearBoard (Ishii et al., 1992), have provided solutions in this respect. However, if one wants to use them in remote support for collaborative work involving manipulation of physical objects, which requires allowing for the objects to be spread out in a three-dimensional space and for participants to move around, they show clear limitations. In these cases, it is necessary to devise systems for supporting remote collaboration between many participants in different positions and different environments (Kuzuoka et al., 1992, 1994, 1995) An essentially flat work-support system like ClearBoard cannot be turned into one that supports collaboration in three-dimensional space; and it is physically impossible to accomplish perfect eye contact if many participants are allowed to move around in the work space.

The purpose of this research is not simply to point out the physical limitations of these CSCW systems, but it is to try to delineate clearly the intrinsic limitations. We refer to the concept underlying in the distance education systems in use today as the 'face-to-face metaphor' (Figure 1), which ultimately aims to create an environment in which distant people talk as if in close face-to-face conversation. However, it is still not clear enough how the face view is used in collaboration, why it is helpful, and what arrangement of audio/visual resources is appropriate for collaboration.

We have performed some remote collaboration experiments with AlgoBlock, a computer tool for collaborative learning, by deploying cameras and monitors in distinct spatial arrangements. As regards the use of several cameras and monitors we were influenced by the work of Gaver and Heath on Multiple Target Video (MTV) (Gaver et al., 1993, Heath et al.,1995) who analyzed which pictures are important and what kind of image people pay attention to. However, our pilot study conducted beforehand revealed

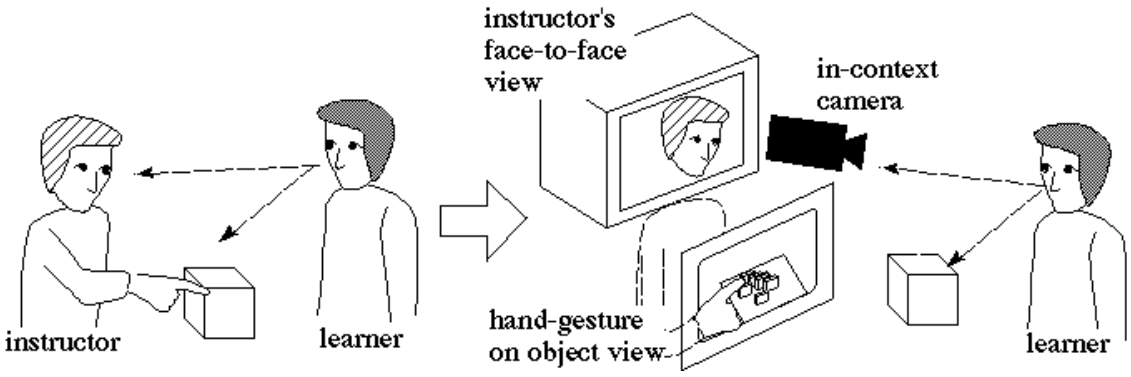

Figure 1. The face-to-face metaphor concept

that many cameras and monitors could often cause user's confusion. Too many resources were not always beneficial, so how to configure those resources has become our concerns. We, therefore, concentrate on the question of how the images on monitors of various parts of the body (face, hands, and so on) should be arranged in the work space. That is, our aim is to find out how people use body images distributed in various ways in a shared work-space as resources for reciprocal actions. Accordingly, we want to establish the validity of redistributing the body images and bodies themselves in a way based on a body metaphor.

Body metaphor concept

From ethnomethodological studies on cooperative work, by Heath (1986), Goodwin (1981), Nishizaka (1991), Yamazaki (1994), and Yamazaki et al. (1996), we know that even more important than the face-to-face view is to make it possible that the instructor can see the learners are looking at what he or she is pointing out and that the learners can become aware of this circumstance. We realize that this reflexive awareness is accomplished through the interaction and positioning of bodies in a shared space.

Based on a pilot study we conducted beforehand, we analyzed how instructors and learners positioned themselves during instruction. When the instructor points at an object, the following points seem to be important: (1) the learners should be able to see the instructor's pointer; (2) the instructor should be able to see that the learners are orienting themselves towards the pointed object as well as the pointer, when they are observing the instructor's pointing; (3) the instructor should be able to reassure the learners by words or actions that the instructor is aware of the learners' orientation; (4) the learners should be able to see the face of the instructor, when they want to know how the instructor is evaluating their behavior or when they want to draw the instructor's attention; and (5) the instructor should be able to notice the learners' orientation toward the instructor himself/herself as well.

Conditions (1) through (5) are achieved naturally, as long as the instructor's pointer is in front of both the learner and the instructor and the learner is in the instructor's field of vision. Here again it is important that the instructor's hand and face are separated enough to make it possible for the instructor to discern which the learner is watching, the hand or the face. We considered how to apply this physical arrangement,

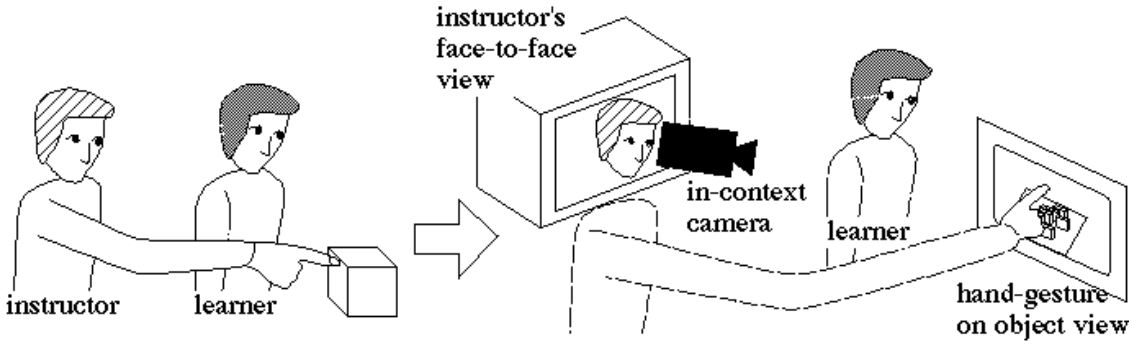

Figure 2 The body metaphor concept

occurring in everyday instruction, to the arrangement of monitors and cameras in a video-mediated communication system. We attempted to use the ordinary body arrangement as a metaphor for the placement of cameras, the face view, and hand-gesture monitors (Figure 2), and named it 'body metaphor'.

Experiments

Task

We used AlgoBlock (Suzuki et al., 1995) as the subject matter for the remote collaboration experiments. AlgoBlock is an educational programming language incorporated into physical blocks (Figure 3). Each command corresponds to a kind of block and the learner can construct a program by connecting blocks to each other by hand. Learners are supposed to build a simple program for guiding a submarine on a CRT screen to its destination.

AlgoBlock was devised as a tool to draw the originally individual work of writing a program into the sphere of collaboration. By giving the commands of the programming language, which are usually hidden inside the computer, a physical form such as blocks, the work of programming is turned into an overt action involving visible body movements. The movements can be observed by other participants. Such an observation can then be reciprocally used as a resource in cooperative work. By transforming a mental work into physical actions, AlgoBlock can enrich the resources of collaboration.

There are two reasons why we have employed AlgoBlock for this study. First, because of the reasons above, AlgoBlock makes it easy to analyze how the collaboration proceeds. That is, the resource used by the learners to achieve the goal of collaboration becomes visible to an outside observer as well as to the learners themselves. Second, the task using AlgoBlock can be a good model of collaborative work involving physical movements, such as in a scientific experiment or in physical training.

Workspace setting

In the experimental workspace of both the instructor's and the learners' sites, the arrangements of communication resources, such as video monitors, video cameras, and AlgoBlocks, were as follows.

The learners' site had three cameras (two in front of the learners and one on the ceiling) and three monitors. One of the front cameras was able to pan, tilt, and zoom according to remote control by the instructor. The other, referred to as the in-context camera, was fixed in such a way as to simultaneously capture the table used for assembling AlgoBlocks, the monitors in the back of the room, and all the learners. The ceiling camera captured all the learners and the whole work space. The three monitors showed the instructor's face, his/her hand gestures, and the AlgoBlock screen. The in-context camera and the remote controlled camera were set in the vicinity of the face-to-face view. The monitor, referred to as the hand-gesture view, showed image from hand-gesture camera in instructor's site, in which instructor's hand-gesture appeared. The instructor's voice was produced from the monitor with the face-to-face view.

The instructor's site had two cameras: one, referred to as the hand-gesture camera, for capturing the instructor's hand gestures on the hand-gesture monitor and the other, referred to as the face-to-face camera, for the face of the instructor. Three monitors were placed in front of the instructor. On the left, from the instructor's point of view, was a personal computer display for remote control of the camera mentioned earlier, with the view taken by that camera. The monitor in the middle, in the vicinity of which the face-to-face camera was set, showed the view of the in-context camera. The monitor to the right, referred to as the hand-gesture monitor, could be switched between the in-context

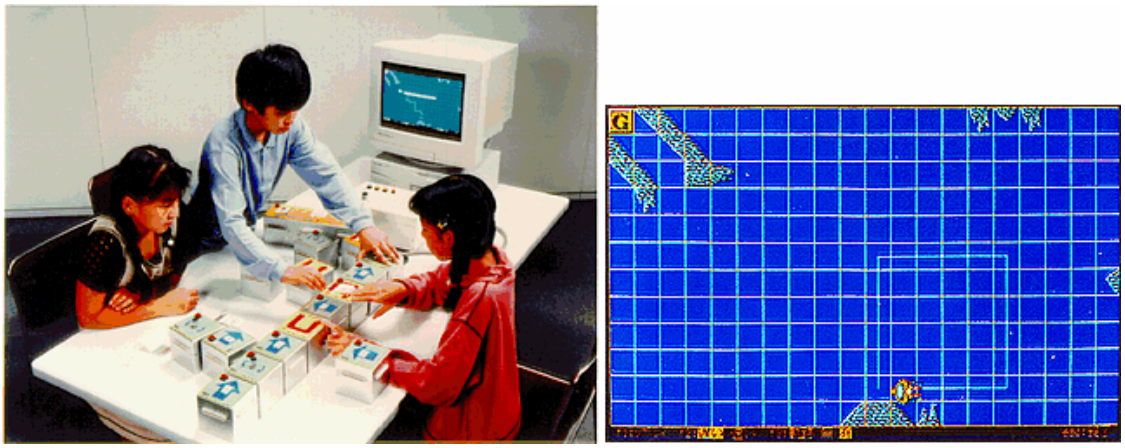

Figure 3. An example of the AlgoBlock session (left) and the screen image (right)

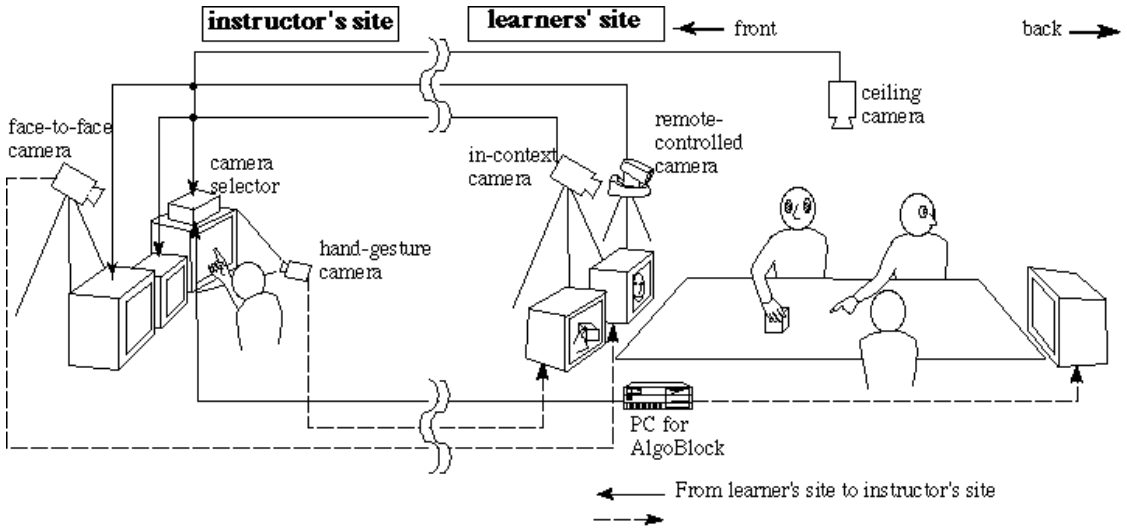

Figure 4. Experimental setting of pattern-2 (body metaphor)

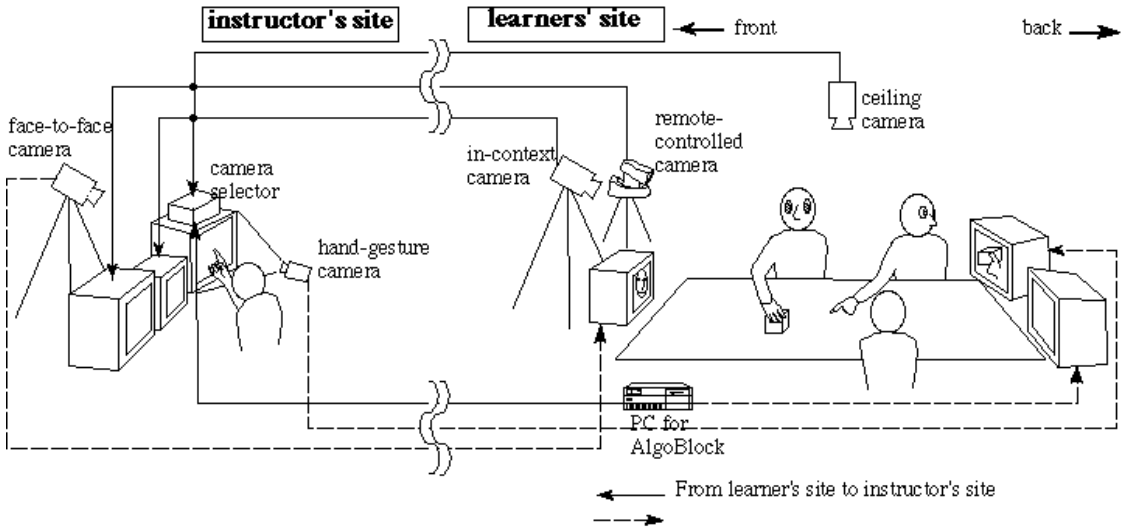

Figure 5. Experimental setting of pattern-1 (face-to-face metaphor)

camera, the remote controlled camera, the ceiling camera, and the AlgoBlock screen. Since this monitor was viewable in the hand-gesture camera, learners could observe the instructor's pointer on something showing on the hand-gesture monitor.

- Pattern-1: This setting was taken after the tele-conference-based distance education systems, i.e., the face-to-face metaphor. The front of the room had the face-to-face view (right) and the hand-gesture view (left). The AlgoBlock screen was placed at the back of the room (Figure 4). In short, all the devices for communicating with the instructor were gathered on one side of the room.

- Pattern-2: This setting was based on the body metaphor. Only the face-to-face view was in front, and the hand-gesture view and the AlgoBlock screen were placed at the back (Figure 5).

		P--------- P----------
I1	HM	----------C-----HM----------C---------HM----------C
	tunagerutokiwa	kono block o konodengen o kitte kirankya ikenain desuyone
	(when you connect these blocks,	this switch ,you have to switch off power.)
L2	AL	----------H----------
		hai
		(yes)
L3	H	----------

Table 1. Transcript from pattern-2 setting experiment

In pattern-2, we paid special attention to the following points:

1) At the learners' site, the hand-gesture view should be positioned opposite the instructor's face-to-face view.

2) The in-context camera should be positioned so that an instructor could see the hand-gesture view in the context of the learner's site.

The only difference between these two settings was the location of the hand-gesture view in the learner's site. It addresses the issues of whether the instructor could see his/her own hand gestures in the context of the learner's site and whether the instructor could clearly determine which of views the learner was looking at. Pattern-2 make it possible for the instructor to check whether the learners were looking at instructor's hand-gesture and how they were responding to the instruction.

Other conditions

The experiment was conducted with four groups: each group consisted of one instructor and two learners. All subjects were university students. An experimental session for each group took one hour in which both patterns were tried. Experimenters taught the instructors how to use AlgoBlock in advance.

Observation

By taking a close look at how the monitors transmitting body images were mutually used as resources for collaboration and communication, we noticed the following.

Observableness of learner's orientation

In pattern-2 when giving directions, instructors used their in-context view to check how learners watched the hand-gesture view (Figure 6).

When the instructor gave particularly detailed

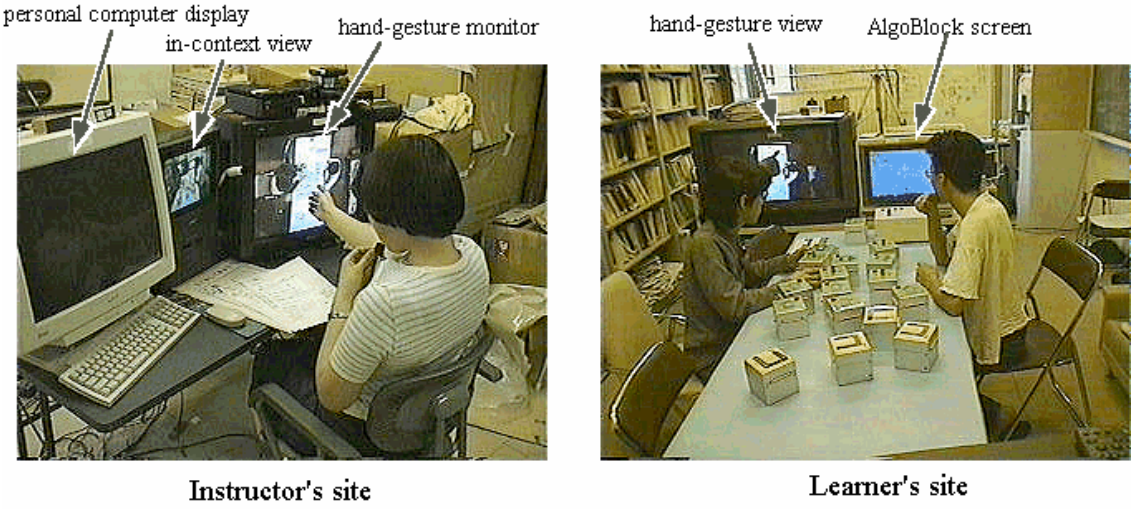

Figure 6. Scenes from the experiment using pattern-2 setting (left: instructor's site, right: learners' site shown in the in-context view). Face-to-face view is next to the camera, and hand-gesture view and AlgoBlock screen are facing the camera.

directions, she alternatively viewed the hand-gesture monitor and the in-context view. At this time the instructor made sure that, while giving instructions, the learner was watching the pointing finger (Table 1).

In this transcript, time proceeds from left to right. Here we had two learners L2 and L3 at the learners' site, and an instructor I1 at the instructor's site. The horizontal rows beginning with the subjects symbols give utterances in Japanese, and the English translations appear in parentheses below. The direction of gaze is recorded above the dialog, where ---- indicates continuation. AL and H stand for the AlgoBlock screen and the hand-gesture view at the learners' site, respectively, HM and C are used for the hand-gesture monitor and the in-context view at the instructor's site, and P indicates that the instructor was pointing at the hand-gesture monitor.

Instructors ascertained, before giving directions, whether or not the learner took a posture of watching the hand-gesture view. When learners did not take up a watching position, instructors used the indexical word 'kono'(these) to get the learners' attention and make them watch the hand-gesture view. Instructors also interrupted instruction when they noticed that learners were not watching the hand-gesture view.

We discovered that, in this kind of setup (pattern-2), when the learner watches the hand-gesture view, the posture indicates 'recipiency' (Goodwin, 1981, Heath, 1986), which is a social display to show that one is ready to receive information. It turned out that through the in-context view in pattern-2 instructors monitored whether learners were showing recipiency for instruction.

In pattern-1 (Figure 7), by comparison, it was difficult for instructors to monitor recipiency. As shown in Table 2, the instructor (I4) asked 'Can you see my hand?' 'Does my hand show?' and even when the learner (L5) affirmed 'Yes, I can.', these questions were often asked repeatedly. Instructors also often repeated instructions concerning the same object using demonstrative pronouns like 'kore wo'(this) and phrases like 'kore wo koko ni'(this here). Moreover, instructors exaggerated their pointing actions on the hand-gesture monitor by moving their hands vigorously.

I4:	<watching in-context view and hand-gesture monitor alternatively>
	At first, can you see my hand?
	<looking at in-context view>
L5:	Ah, Yes, I can.
I4:	Oh, you can.
I4:	These blocks can be connected both in row and in column.
L5:	Yes.
I4:	And, what can I say, this submarine, here can you see a submarine, can't you?
L5:	Yes.
I4:	This, in the first place.
L5:	Ah, Yes.
I4:	This is a goal, a goal.
L4:	Well, does my hand show?

Table 2. Transcript from pattern-1 setting experiment

The cause of the difficulty was that instructors could not monitor how learners were watching their pointing. Because the hand-gesture view was located beside the in-context camera and near the face-to-face view, instructors were not able to know whether or not learners were showing recipiency. In this way, pattern-1 caused communicative asymmetry (Heath et al., 1991, 1992), which occurs in video-mediated communication.

Awareness of an instructor

When in pattern-2 the instructor moved or placed his/her finger continuously on the hand-gesture monitor, it was utilized as a reciprocal resource for interaction. As shown in the transcript (Table 1), when the learner (L2) looked at the hand-gesture monitor, and said 'hai (yes)', the instructor (I1) removed her finger and started to point at another object. The movement of the instructor's finger showed that she understood the learner's action.

It can be said that the instructor's finger movement becomes a reciprocal instrument indicating that the instructor was monitoring the learner's actions. This, coupled with the instructor's face-to-face views and the instructor's oral response, constitutes a valuable resource for the learner to monitor that the instructor is watching the learner's action. In other words, it can enhance the learner's awareness of the instructor.

In contrast to this, in pattern-1, the instructor sometimes removed his/her finger from the monitor before the learners displayed their understanding of what the instructor was pointing to, or kept it on the screen for an unnecessarily long time. These were because the instructor was not aware of whether or not, or how carefully, the learners were watching his/her finger. Conversely, when the learners tried to ascertain the instructor's response by looking at the hand-gesture and the face-to-face views, this did not always go smoothly either, since the instructor did not know which of the views (hand-gesture or face-to-face) the learners were looking at.

Surrogate for an instructor

When, in pattern-2, learners tried to ask a question, or when the AlgoBlock operation had been successfully completed, they often looked at the face view to check the instructor's reaction and approval.

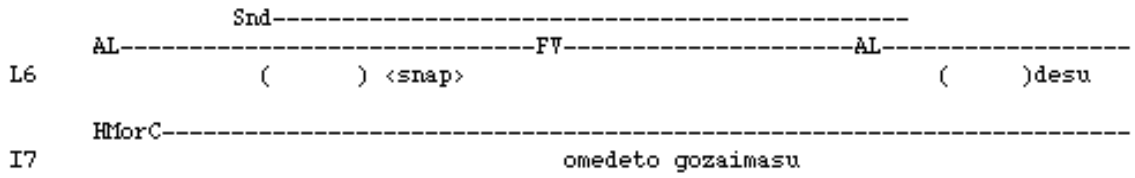

Table 3. Transcript from pattern-2 setting experiment

Table 3 shows a transcript from pattern-2 when the learner (L6) succeeded in reaching a goal. As shown in the transcript, the instructor (I7) was looking at either the hand-gesture monitor (HM) or the in-context view (C), but which view they were watching was not clear in the videotape data, and the learner was looking at the AlgoBlock screen (AL) to check the result of his program execution. As soon as he realized that the trial was successful (when the sound effect representing the submarine's arrival to the goal (Snd) started), he said something that sounded like an expression of success and snapped his fingers in both hands. Then, he turned his face to the face-to-face view (FV). At that moment, the instructor said "Congratulations!." In this way, the boundary of a series of activities was marked interactively by those actions.

There are several relevant timings when the instructor could give congratulations: at the moment the submarine arrives at the goal, after the learner's remark, or after the sound effect. As well as these, this observation proved that the learner's action of turning his face to the instructor's face view could also provoke the instructor's reaction. In this sense, it can be said that the instructor's face view in the learner's site of pattern-2 played the role of the instructor's representative, or surrogate. This happened because in pattern-2 the face view was far enough apart from the other resources for the instructor to notice the learner's orientation to it. In this fragment, by using the face-to-face view as a surrogate, participants interactively marked the activity's boundary, and thus to achieve the end of a session successfully.

In pattern-1, by contrast, reply, reaction, and evaluation were not always immediately performed. The instructor, not knowing learner's orientation to himself/herself, often failed to catch the proper time to intervene in a learner's activity.

Remaining problems

However, there were still some problems common to both patterns.

One of the problems was that learners could not know which views the instructor was oriented towards. There was an asymmetry in communication resources between learners and the instructor, i.e., learners could not observe the instructor's in-context view, while the instructor could do. Therefore, it sometimes happened that learners began to talk regardless of instructor's state of attention, although the trouble was restored easily by conversation. Learners tended to anticipate that the instructor was watching their actions at any time, while this was not always the case.

This asymmetry in system design arises from the difference in role between an instructor and a learner, since we presupposed an educational situation rather than equally collaborative work.

This problem may be solved by introducing a new in-context view at the learner's site just like at the instructor's site. However, adding more monitors may possibly not be as effective as we suppose, judging by the following problem.

Another problem was the instructor's confusion in utilizing communication resources; the instructor sometimes pointed outside the hand-gesture camera's capture range on the hand-gesture monitor, or even pointed at a monitor that learners could not see, and the same kind of confusions may possibly happen also at the learner's site. The reason was that it was not always clear what information was transferred to the other site, although it was possible for the instructor to know, if he/she carefully watched the in-context view. However, it could be troublesome for the instructor to have to watch another monitor in order to see how his/her own fingertip showed on the hand-gesture view in the learner's site.

A possibility is that there were too many resources to utilize, or that the relationship among various communication resources was too complicated to understand. Therefore, closer examination to validate the necessity of individual resources and/or more sophisticated integration of those resources needs to be investigated in our future work.

Conclusions

It is well known that in everyday communication the relative position of person and person, or person and machine, plays an important role (Goodwin, 1981, 1995). One may assume that since communication takes place using the bodies of individuals, it depends strongly on the deployment of those bodies. If one wants collaboration through a video-mediated communication system to occur in a natural way,

spatial arrangement of functional resources originating from body part functions becomes a fundamental issue. Consequently, we have proposed the concept of a body metaphor in configuring a video-mediated communication system.

By deploying three communication resources (objects for manipulation, face view, and hand-gesture view) on the basis of the body metaphor, it became clear which of them learners were orienting to. Specifically, the instructor could see learners looking at the instructor's pointing. Hence, the instructor could recognize the learner's recipiency, and learners were aware of instructor's observation. In addition, it turned out that the learner could use the face view as a surrogate of the instructor to call the instructor's attention and to interactively mark activity boundaries.

We have been working on the field study of a classroom at school and a workplace of emergency care for several years. Taking such findings into account in addition to the findings discussed here, we will try to refine the body-metaphor concept by figuring out a better balance between understandability and richness of communication resources, so that we can develop more practical systems for remote collaborative learning or remote medical procedures.

Acknowledgments

The authors would like to thank Dr. Tammo Reisewitz and Mr. Ron Korenaga for their contributions to this work. We also thank Dr. Graham Button for his insightful comment on this paper. This research was supported by grants from Grant-in-Aid for Scientific Research (A) 07309018 (Head Investigator: Keiichi Yamazaki), 1996 CASIO Science Promotion Foundation, and 1996 the Telecomunication Advancement Foundation (TAF).

References

Gaver, W., Sellen, A., Heath, C., and Luff, P. (1993): "One is not enough: Multiple Views in a Media space", Proc. of INTERCHI'93, 1993, pp. 335-341.

Goodwin, C. (1981): Conversational Organization: Interaction between speakers and hearers, Academic Press, New York.

Goodwin, C. (1995): "Seeing in Depth", Social Studies of Science 25, pp. 237-274.

Heath, C. (1986): Body Movement and Speech in Medical Interaction, Cambridge University Press, Cambridge.

Heath, C., and Luff, P. (1991): "Disembodied Conduct: Communication through video in a multi-media environment", Proc. of CHI'91, 1991, New Orleans, pp. 99-103.

Heath, C., and Luff, P. (1992): "Media space and communicative asymmetries: preliminary observation of video-mediated interaction", Human Computer Interaction, 7(3), pp. 315-346.

Heath, C., Luff, P., and Sellen, A. (1995): "Reconsidering the Virtual Workplace: Flexible Support for Collaborative Activity", Proc. of ECSCW'95, 1995, pp. 83-99.

Ishii, H. and Kobayashi, M. (1992): "ClearBoard: A Seamless Medium for Shared Drawing and Conversation with Eye Contact", Proc. of CHI'92, 1992, pp. 525-532.

Kuzuoka, H. (1992): "Spatial Workspace Collaboration: SharedView Video Supported System for Remote Collaboration Capability", Proc. of CHI'92, 1992, pp. 33-42.

Kuzuoka, H., Kosuge, T., and Tanaka, M. (1994): "GestureCam: A Video Communication System for Sympathetic Remote Collaboration", Proc. of CSCW'94, 1994, pp. 35-43.

Kuzuoka, H. Ishimoda, G., Nishimura, Y., Suzuki, R., and Kondo, K. (1995): "Can the GestureCam be a Surrogate?", Proc. of ECSCW'95, 1995, pp. 181-196.

Nishizaka, A. (1991): The Social Order of Therapy II, Mejigakuin Ronsou, 475, Meijigakuin University, (in Japanese).

Suzuki, H. and Kato, H. (1995): "Interaction-Level Support for Collaborative Learning: AlgoBlock-An Open Programming Language", Proc. of CSCL'95, 1995, pp. 349-355.

Tang, J., and Minneman, S. (1990): "VideoDraw: A Video Interface for Collaborative Drawing", Proc. of CHI'90, 1990, pp. 313-320.

Yamazaki, K. (1994): The Pitfalls of a Beautiful Face--Ethnomethodological Studies on Sexuality, Harvest-sha. Tokyo, (in Japanese).

Yamazaki, K., and Yamazaki, A. (1996): "Ethnomethodology of discrimination--Organization of Category and Organization of Situation", in Series: Contemporary Sociology 15, Iwanami-shoten, Tokyo, (in Japanese).

Author's Address

Hiroshi Kato and Hideyuki Suzuki: C&C Media Research Laboratories, NEC Corporation, 1-1, Miyazaki 4-chome, Miyamae-ku, Kawasaki, Kanagawa, 216, Japan. kato@ccm.cl.nec.co.jp, hideyuki_suzuki@ccm.cl.nec.co.jp.
Keiichi Yamazaki: Faculty of Liberal arts, Saitama University, 255 Shimo-okubo, Urawa, Saitama 338, Japan. yamakei@sacs.sv.saitama-u.ac.jp.
Hideaki Kuzuoka: Institute of Engineering Mechanics, University of Tsukuba, 1-1-1, Tennoudai, Tsukuba, Ibaraki, 305, Japan. kuzuoka@kuzuoka-lab.esys.tsukuba.ac.jp.
Hiroyuki Miki: Media Laboratories, Oki Electric Industry Co., Ltd., 550-5, Higashiasakawa-cho, Hachioji, Tokyo, 193, Japan. hmiki@hlabs.oki.co.jp.
Akiko Yamazaki: 4-3, Haruno 1-chome, Oomiya, Saitama, 330, Japan.