ICAT 2003
December 3-5, Tokyo, JAPAN

Design Suggestions on Remote Pointing in Distant Collaborations

Tomoko Imai, Dairoku Sekiguchi, Naoki Kawakami, Susumu Tachi

The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656 Japan
{timai, dairoku, kawakami, tachi}@star.t.u-tokyo.ac.jp

Abstract

In this paper, we will discuss design suggestions to realize remote pointing in distant collaborations. We use the pointing to anchor what we say to real worlds in our daily life. In the virtual world, we anchor our conversation to virtual worlds. The anchoring is necessary to establish mutual understanding between participants but important elements for the anchoring are not known. We show that the pointing toward a desk space can be done using either eyes or using both the eyes and a hand without significant difference in a pointing accuracy under the face-to-face condition. Also, participants judge pointed locations mainly using a hand cue even if the pointer pointed with her eyes and her hand. Therefore, avatars should be able to show hand information accurately but the relation between the eyes and a hand does not require the accurate representation as the hand information. Stereoscopic images can recreate 73% of information that is provided by a face-to-face pointing and more accurate recreation of the face-to-face pointing requires fine-tunings of the system to each user. We show that it is possible to realize the high pointing accuracy without the system tuning to each user, if we use a simple rod as a remote virtual finger.

Key words: pointing, videoconference, gaze, gesture, remote collaboration

1. Introduction

People often point objects to anchor what they say to real world objects [1]. The anchor can be established by orienting a finger, directing gazes to the objects, or holding them. Understanding the anchor has been known as a crucial element for infants to increase their vocabulary and infants over one year can understand the relation between pointing gestures and pointed targets [2].

Gaze toward other's face has been studied for a long time because it has been regarded as an important element for the social life [3]. Gibson et al. measured relation between gaze directions and a sense of being looked at. They reported that participants felt that they were looked at when a looker looked at a face area that is about 18 cm when seen from 200 cm. Anstis et al. showed that people can perceive other's gaze that is presented using a TV monitor with monoscopic images at an accuracy about 0.96 deg that is less that 4 cm if the distance between the display and a participant is 200 cm. These results suggest that information on whether a person at a remote place is watching a partner's face or not can be transmitted using a monoscopic image. We extend the experiment to measure a perception of gazes toward a work space and show that people perceive other's gaze toward the desk space about the accuracy of 8 cm under a face-to-face condition and perception of the depth component of the gaze direction is difficult. Recently, the gaze direction is measured to design natural or efficient input devices [4]. However, how people understand other's gaze direction is sill under investigation. Since the important cues for the gaze perception is not known, many different approaches has been tried to transmit gaze information to remote places. Hydra system represented gaze directions using a small real object with an image of a remote participant and was compared to different video systems in terms of objective measures of on-off patterns of speech but the difference was not measured [5]. In the virtual reality (VR) applications, participants are represented as 2D or 3D avatars that are controlled by motion data of the users. However, it is not known how 3D information is important to understand other's gaze direction.

While gaze has been studied for a long time and used in many applications, little is known about pointing gestures. Miyasato et al. showed that when a pointer points to one of targets that are placed at 18cm intervals in the horizontal direction and 20cm in the depth direction, participants could tell pointed targets at an accuracy of 71%. When the pointing gestures are presented by a monoscopic image, the accuracy was reduced to 30%. This research result suggests that, pointing capability of a monoscopic image is less than 50% of that for the face-to-face pointing. Pointing applications are usually implemented with remote robots or cursors on screens and there is little applications that try to convey pointing gestures using image [6, 7, 8, 9]. Usually design decision on remote pointing is made

intuitively but knowledge on how people understand other's pointing gestures is necessary to design efficient remote pointing. Therefore, we measure accuracy of human perception of pointing gestures to understand how people understand other's pointing.

In this paper, we discuss design suggestions on efficient remote pointing systems based on data that were gathered in experiments on human perception of pointed location. We consider a monoscopic videoconference system as a basic structure for remote pointing systems and discuss efficient improvement of the system. Firstly, we examine perception accuracy of different pointing method under a face-to-face condition to find out cues that enable efficient pointing. Then, we select appropriate presentation methods for the cues and measure how the pointing information is deteriorated with the approach. Finally, we present design suggestions for remote pointing systems. . The discussion is applicable to VR applications because we study pointing capability of monoscopic, stereoscopic images, and mixture of image and a real object, which are elements of VR technology. Also, data on the face-to-face pointing can contribute to designing new VR applications.

2. Comparison of different pointing methods under face-to-face conditions

To design natural and accurate remote pointing, it is necessary to find out cues that can convey pointed locations efficiently under a face-to-face condition. If there are cues that can show pointed location with higher accuracy than others, representing them with collaboration systems is an efficient and effective approach for a system design. Therefore, we compared the accuracy of perceptions of pointed locations under three different face-to-face conditions; perceiving pointed location by watching both eyes and a hand (the eyes-hand condition), only eyes (the eyes condition), and only a hand (the hand condition), using the experimental environment shown in Fig. 1. Under the eyes-hand condition, participants judged pointed locations by watching both eyes and a hand of the pointer. This condition is the most natural pointing method of the three conditions and expected to be the most efficient pointing method. A participant and a pointer were seated across a table and the pointer pointed targets on the table as shown in Fig 1. The distance between the pointer and the participants was 130 cm and the height of their eye positions was fixed to 114 cm before starting the experiment. The height of the table was 71 cm. The pointer was a Japanese woman who pointed at the target board from a window of a room whose inside lighting conditions were controlled.

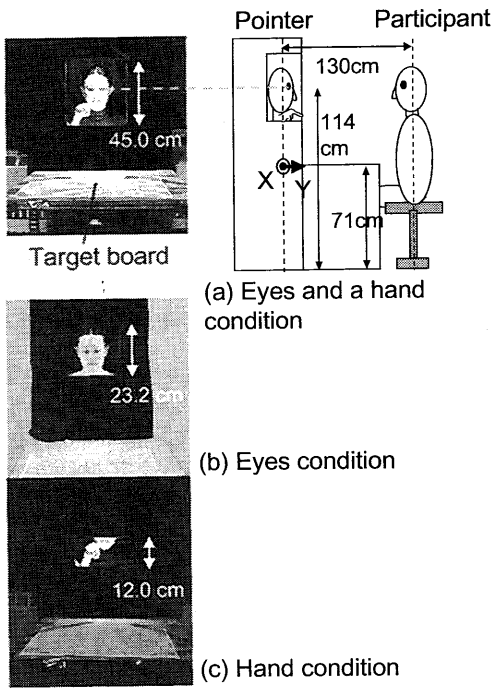

Fig. 1 Experimental environments for the face-to-face conditions

Participants were one Japanese woman and two Japanese men, all in their 20s and familiar with the looker. All of them had normal or corrected-to-normal vision and could see stereo image correctly when their stereo vision was checked with a stereo test (STEREO TESTS, STEREO OPTICAL CO., INC.). The size of the window was 30.7 cm x 45.0 cm for the eyes-hand condition, 30.7 cm x 23.2 cm for the eyes condition, 30.7 cm x 12.0 cm for the hand condition. Under the eyes-hand condition, the pointer placed her fingertip on a line between her right eye to the target. Placing a fingertip on a line between one of the eyes and a target is a reported as a common for human natural pointing but how people point targets are still under investigation [9]. In this experiment, the pointer pointed 100 targets that were chosen in a random order. The targets were small black dots aligned in a grid at intervals of 1 cm. The grid had 54 dots in the horizontal direction (x direction) and 40 dots in the depth direction (y direction), for a total of 2160 dots. The distance from the eye position of the pointer to the closest dot to her was 47 cm.

While she prepared the pointing gesture, the participants closed their eyes. When the pointer fixed her pointing gesture, she said, "Ok" to participants. Then the participants placed a pin on the target board to mark the points that they thought she was pointing at. After placing the pin, they closed their eyes. The pins were numbered from 1 to 100 to show the relation between pins and pointed targets. The pointer pointed at one of

the 2160 points in accordance with a prearranged random order.

Here, we define an **error vector** as a vector drawn from a pointed location and a place that a participant placed a pin. We call the absolute value of the error vector as an **error**. Fig. 2 shows the errors averaged over 100 errors for each participant. Contrary to our expectation, the accuracy of perceived location for the eyes-hand condition and the eyes conditions did not have statistically significant difference as shown in Fig. 2. The error averaged over all the participant for the eyes condition is 8.1 cm and the error for the eye-hand condition is 7.2 cm and slight statistical difference is found ($t(599)=2.43$, $0.05>p>0.01$). The performance under the hand condition was 9.6 cm and is lower than the other two conditions. T-test shows that there are statistically significant differences.

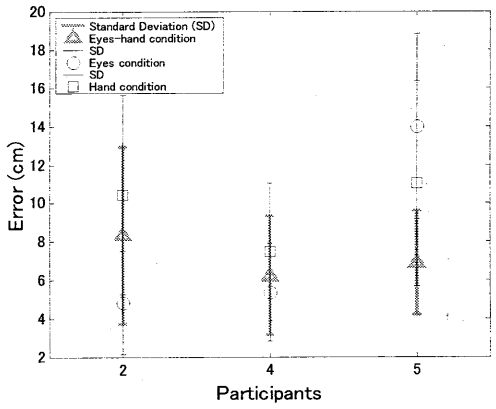

Fig. 2 Errors obtained under Eyes, Eyes-hand, and Hand conditions

This result suggests that a remote pointing can be designed either using the eyes information or the eyes and the hand information. Therefore, if a system is designed to recreate accurate human gesture information, the system can convey pointing information either by eyes or both eyes and a hand. This is the case for transmitting real image of the pointer or controlling a remote robot with the human motion data. However, if a system is designed to represent essences of pointing information, this conclusion cannot be applied. In the next section, we will discuss design suggestions for systems that are designed to transmit essence of the remote participants.

3. Design suggestions for systems that transmit essences of the remote participants.

When a remote participant is represented by CG avatar or a robot that is not controlled by motion data of the participant, the system can modify participant's gesture so that pointing information can be transmitted efficiently. In this case, information on the cues that are important for perception of pointed location is necessary. In our preliminary studies, we found that people try to find a pointed location by extrapolating an orientation of a finger even if they can watch both eyes and a hand at the same time. The perceived results under hand condition were not as accurate as the data obtained under eye-hand condition because a pointer was not able to orient her finger to the target location correctly and the major reason for the lowest performance of participants under the hand condition was the misleading finger orientation.

If the pointer is represented by a CG or a robot that are not controlled by the pointer's motion data, it is possible to orient a virtual finger to the target correctly. When we orient a virtual finger that is shown in Fig. 3 to a target, participants perceived the pointed location much accurate than under the hand conditions. The experimental environment is the same as the one for the face-to-face condition as shown in Fig. 4 but real person was replaced by a virtual figure.

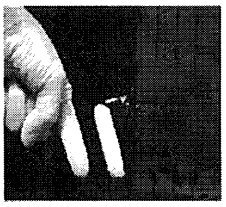

Fig. 3 A real finger and a virtual finger

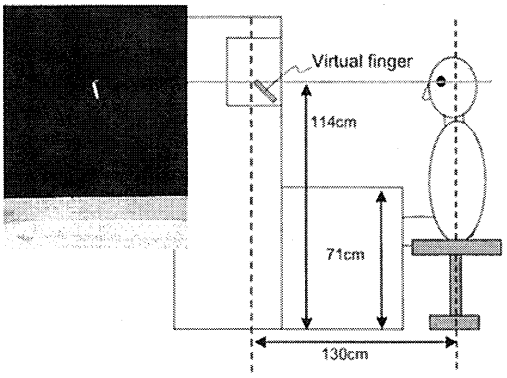

Fig. 4 Experimental environment for the virtual finger condition

As a preliminary study, we measured the perception accuracy of three participants; participant2, 4, and 5 as shown in Fig. 5. Participants4 did not show statistically significant difference between the eyes-hand and the virtual finger condition ($t_4(99)=1.51$, $p_4>0.05$). Compared to the hand condition, participant2 and 4 performed better under the virtual finger condition ($t_2(99)=1.66$, $p_2<0.05$, $t_4(99)=1.66$, $p_4<0.05$). These

results suggest that there is a possibility that a face-to-face like pointing condition might be recreated by orienting the virtual finger correctly to the target.

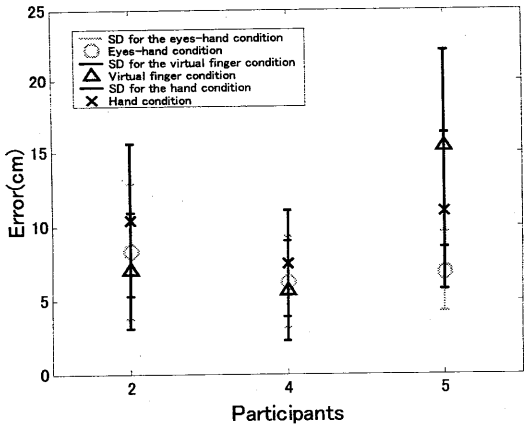

Fig. 5 Comparison of accuracy of perceived result

When we define a **pointing capability** of the virtual finger as

$$\frac{ERROR_{EyesHand}}{ERROR_{VirtualFinger}} \times 100 \qquad (1)$$

$ERROR_{VirtualFinger}$: the average of the absolute values of the error vector for the virtual finger condition

$ERROR_{EyesHand}$: the average of the absolute values of the error vector for the eyes-hand condition

, the virtual finger recreated about 77% of the pointing presented under face-to-face condition. This result shows that the virtual finger cannot provide the same quality of the pointing information as the pointing with both eyes and a hand under the face-to-face condition. However, the virtual finger can recreate the face-to-face condition better than the monoscopic image because Miyasato et al. showed that presenting pointing gestures with monoscopic image can convey less than 50% of the information that is presented under face-to-face condition [6]. Therefore, using the virtual finger to represent pointing is an efficient human interface design than using a monoscopic image. Our results support the intuition of researchers who used rods to realize remote pointing rather than using images.

4. Design suggestions for systems that transmit accurate information of the remote participant.

When a remote surrogate represents essences of the remote participant, the virtual finger can be used to convey pointed location at remote sites. However, when the remote surrogate is designed to represent exact remote participant's gesture, the virtual finger cannot be used. Here, the surrogate can be a robot controlled by data of a remote participant or images of the participant.

In this experiment, we examine transmitting pointing information with eyes using images because our experiments showed that a pointing can be done using either eyes or eyes and a hand without significant difference in the pointing capability. Also, it has been showed that representing the finger orientation with image is not efficient. Pointing capability of robots that can recreate participants' eye expressions should also be measured but it is not possible to develop the system using currently technology.

We compared transmitting gaze directions using the monoscopic and stereoscopic images using experimental environments shown in Fig. 6.

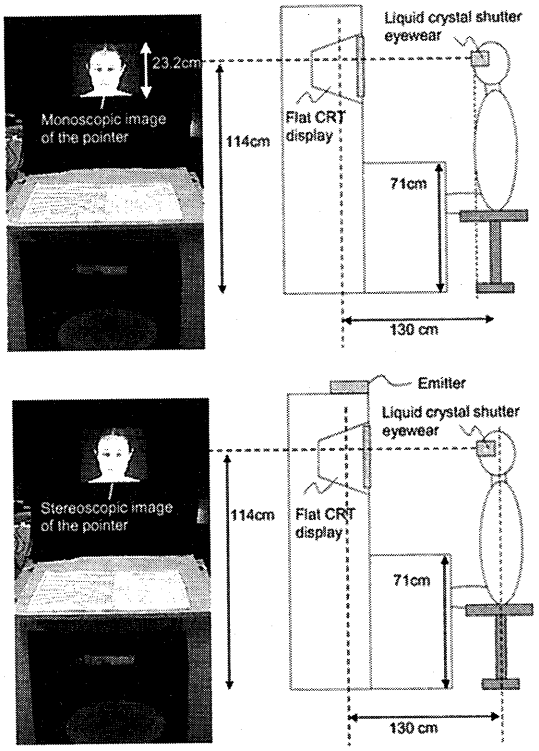

Fig. 6 Experimental environments for the monoscopic and the stereoscopic image experiments

The participants wore the liquid crystal shutter eyewear under both the monoscopic and the stereoscopic image conditions to eliminate influences of the eyewear. The monoscopic images were taken by a camera that was placed at a position that is the same as the center of eyes of a participant. The stereoscopic images were taken by two cameras whose distance was 6.5 cm that is an average of the pupillary distance of the participants as shown in Table 1. The pointer was the same as for the

eyes-hand experiment. The same three participants for the eyes-hand condition experiment and additional three participants judged pointed locations in this experiment.

Table 1. Pupillary Distances

Participant	P1	P2	P3	P4	P5	P6
PD(cm)	6.8	6.7	6.1	6.4	6.6	6.2

The average of errors for all the participants under the face-to-face condition, the stereoscopic condition, and the monoscopic condition were 8.0, 11, and 12 cm. T-test showed that there is statistically significant difference between the monoscopic and the stereoscopic conditions ($t(599)=4.66$, $p<0.01$). This result suggests that stereoscopic image improved the pointing capability of the monoscopic gaze image. However, there is still significant difference between the stereoscopic and the face-to-face conditions ($t=(599)=10.33$, $p<0.01$). Therefore, the stereoscopic image was not able to recreate the face-to-face condition. It can be said that the monoscopic image can recreate 67% of the face-to-face condition and the stereoscopic image can recreate 73% of the face-to-face condition using the equation (2).

$$\frac{ERROR_{face-to-face}}{ERROR_{stereo}} \times 100 \qquad (2)$$

$ERROR_{face-to-face}$: the average of the absolute values of the error vector for the face-to-face condition

$ERROR_{stereo}$: the average of the absolute values of the error vector for stereoscopic image condition

When we examined the data for each participant, only participant2 shows statistically significant improvement under the stereoscopic image condition over the monoscopic image condition ($t(99)=5.56$, $p<0.01$) as shown in Fig. 7. The participant1 did not show significant difference between the face-to-face and the stereoscopic image condition ($t(99)=1.23$, $p>0.05$). Other participants performed poorer under the stereoscopic image condition than under the face-to-face condition. Comparison to the pupillary distance of the each participant revealed that participant3 who has the narrowest pupillary distance had the largest difference between the result obtained under the face-to-face and the stereoscopic image conditions. Participant2 who has the next to the narrowest pupillary distance performed better under the stereoscopic image condition than under the face-to-face condition. This result suggests that participnat2 did not receive correct stereo information from the images. We suppose that the perception accuracy of the gaze direction is influenced by the difference between the stereo camera distance and the participant's pupillary distance. Our preliminary studies suggest that adjusting the stereo camera distance to each participant's pupillary distance and presenting image according to the participant's head position may contribute more than 10% increase in recreating the face-to-face condition.

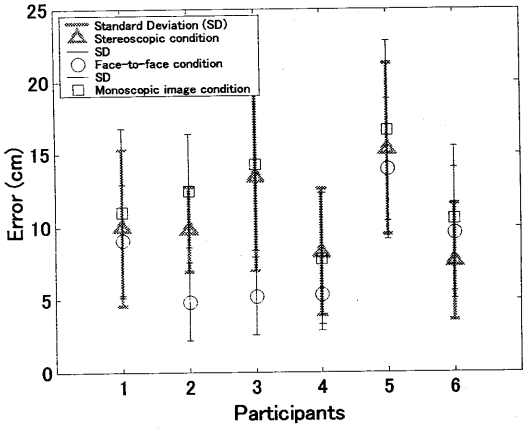

Fig. 7 Errors for the face-to-face, the stereoscopic, and the monoscopic image conditions

5. Comparison on Different Approaches for Remote Pointing

Under the face-to-face condition, the errors averaged over all the participants did not show significant difference between the eyes and the eyes-hand conditions. However, when a target is pointed using both eyes and a hand, participants judged pointed location mainly using the hand information and the virtual finger experiment showed that the virtual finger can present essence of the hand information. Therefore, systems can realize the remote pointing either using real images of the face of a remote participant or using the virtual finger. When we compared the pointing capability of the monoscopic image condition, stereoscopic condition, and the virtual finger condition as shown in Fig. 8, the participants performed best under the virtual finger condition. The average of errors taken over participants2, 4 and 5 for the stereoscopic image condition is 11 cm and the error for the virtual finger condition is 9.3 cm and this difference is statistically significant ($t(299)=4.87$, $p<0.01$).

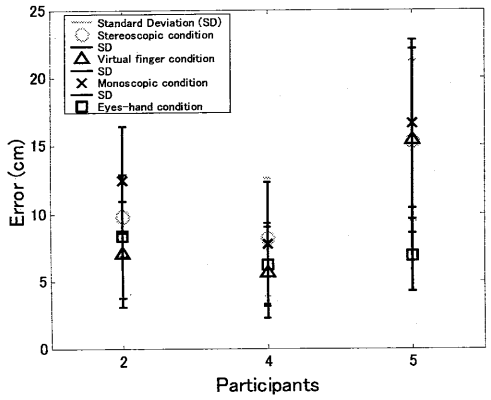

Fig. 8 Errors obtained with different pointing methods

If we try to design systems that can convey much accurate pointing information than the virtual finger using images, stereo cameras need to be adjusted to each viewer's pupillary distance and his/her head needs to be tracked. Also, we reported that people switch cues that are used to understand conversations. While an avatar talked without pointing gesture, people looked into the avatar's face but she only watched its hand when it started pointing [11]. Therefore, it might cost-effective to switch mode of communication systems according to the conversation mode. When a conversation needs not to point real objects, participants use image that show other's face. If they need to point to real objects, they can switch to use virtual fingers.

6. Conclusion

In this paper, we discussed how to realize the remote pointing that is similar to the face-to-face pointing with a cost-effective design based on human perception measurements. We showed that the pointing toward a desk space can be done using either eyes or using both the eyes and a hand without significant difference in a pointing accuracy under the face-to-face condition. Also, participants mainly watched her hand even if the pointer pointed with her eyes and her hand. The stereoscopic image was not able to recreate the same pointing accuracy and our preliminary tests suggest that the recreation require fine-tuning of the system to each user. However, we showed that it is possible to realize the high pointing accuracy without the system tuning to each user, by using a simple rod as a remote virtual finger. These results suggest that adding the virtual finger to a monoscopic video conferencing system is a cost-effective improvement of the communication system to enable remote pointing because if we try to transmit the pointing information with the same accuracy using only images, elaborate tuning of the system to each participant is required. Also, avatars should be able to show hand information accurately but the relation between the eyes and a hand does not require the accurate representation as the hand information.

Acknowledgements

The research has been supported by Japan Science and Technology Agency (JST) and conducted as a part of the Core Research for Evolutional Science & Technology Project by JST.

References

1. S. Kita: "Pointing: Where Language, Culture, and Cognition Meet," *Lawrence Erlbaun Associates, Publishers, NJ*, USA (2003).

2. M. Scaife and J. S. Bruner: "The capacity for joint visual attention in the infant," *Nature*, 253 (24), pp. 265-266 (1975).

3. S. Baron-Cohen: "Mindblindness," *Cambridge, MA, MIT Press.* (1995).

4. D. J. Ward and D. J. C. MacKay: "Fast hands-free writing by gaze direction," *Nature*, 418 (22), p. 838 (2002).

5. A. J. Sellen: "Speech patterns in video-mediated conversations," *Proc. of CHI'92*, pp. 49-59 (1992).

6. T. Miyasato and F. Kishino: "An Evaluation of Virtual Space Teleconferenceing System Based on Detection of Objects Pointed through a Virtual Space," IEICE Transactions on Information Systems, Vol. J80-D-II, No. 5, pp.1221-1230 (1997).

7. S. R. Fussell, L. D. Setlock, E. M. Parker, J. Yang: "Assessing the Value of a Cursor Pointing Device for Remote Collaboration on Physical Tasks," *Proc. of CHI 2003* (2003).

8. H. Kuzuoka, S. Oyama, K. Yamazaki, K. Suzuki, and M. Mitsuishi: "GestureMan: A Mobile Robot that Embodies a Remote Instructor's Actions, "*Proc. of CSCW'00*, pp. 155-162 (2000).

9. E. Paulos and J. Canny: "ProP: Personal Roving Presence," *Proc. Of CHI'98*, pp. 296-303 (1998).

10. Henriques, D. Y. P. & Crawford, J. D., Role of eye, head, and shoulder geometry in the planning of accurate arm movements. *J Neurophysiol* , 2002. **87**:p. 1677-1685.

11. T. Imai, A. Johnson, J. Leigh, D. Pape, T. DeFanti, "Supporting Transoceanic Collaborations in Virtual Environment", *Proc. of APCC/OECC'99, Beijing China*, pp. 1059-1062 (1999).